Break it, Imitate it, Fix it: Robustness by Generating Human-Like Attacks

Anonymous authors
Paper under double-blind review

Abstract

Real-world natural language processing systems need to be robust to human adversaries. Collecting examples of human adversaries for training is an effective but expensive solution. On the other hand, training on synthetic attacks with small perturbations—such as word-substitution—does not actually improve robustness to human adversaries. In this paper, we propose an adversarial training framework that uses limited human adversarial examples to generate more useful adversarial examples at scale. We demonstrate the advantages of this system on the ANLI and hate speech detection benchmark datasets—both collected via an iterative, adversarial human-and-model-in-the-loop procedure. Compared to training only on observed human attacks, also training on our synthetic adversarial examples improves model robustness to future rounds. In ANLI, we see accuracy gains on the current set of attacks ($44.1\% \rightarrow 50.1\%$) and on two future unseen rounds of human generated attacks ($32.5\% \rightarrow 43.4\%$, and $29.4\% \rightarrow 40.2\%$). In hate speech detection, we see AUC gains on current attacks ($0.76 \rightarrow 0.84$) and a future round ($0.77 \rightarrow 0.79$). Attacks from methods that do not learn the distribution of existing human adversaries, meanwhile, degrade robustness.

1 Introduction

Improving accuracy on real adversarial examples is critical to effective natural language processing (NLP) systems. In this paper, we propose methods to improve adversarial robustness beyond what we can achieve by either directly training on real adversarial examples or on simple synthetic attacks. We achieve this by learning the distribution of real attacks, and generating synthetic examples that imitate them.

Adversarial examples were designed by humans to induce misclassification (Szegedy et al., 2014). Adversarial robustness is the ability of a classifier to correctly label adversarial examples. Adversarial robustness has been extensively studied on benchmark NLP tasks (Jia & Liang, 2017; Ettinger et al., 2017; Zhang et al., 2020; Gao et al., 2018). Despite extensive work and progress in this domain, NLP classifiers are still not robust to real-life text adversaries (Lees et al., 2021; Borkan et al., 2019). These failures can result in real-world harms (Scheuerman et al., 2021). Prior work relies on two types of approaches: training on adversarial examples collected by humans (Dinan et al., 2019; Nie et al., 2020), and synthetic attacks (Uesato et al., 2018; Jin et al., 2020). Vulnerability in real-world scenarios is, in part, due to (a) too few real adversarial examples available, and (b) over-simplification of synthetic attacks to produce adversarial examples at scale.

Collecting human-generated NLP adversarial datasets is expensive (Xu et al., 2021; Hendrycks et al., 2021) — even more prohibitively so for modern large models (Ganguli et al., 2022a). Thus, we rely on synthetic examples to improve robustness beyond what we can achieve by directly training on our limited number of human-generated examples. Existing synthetic attacks, however, are often over-simplified; because, unlike computer vision, it is difficult to generate synthetic examples with the desired class label in a discrete space. So these attacks are often reduced to easy-to-generate attacks based on easy-to-compute text-quality metrics (Uesato et al., 2018).

Popular synthetic NLP attacks include (a) template-based or small text-edit-distance attacks, (Malfa & Kwiatkowska, 2022; McCoy et al., 2019b; Zang et al., 2020; Ren et al., 2019), (b) perturbation attacks that use word embeddings and search within an ε-neighborhood (Jia et al., 2019a; Zhao et al., 2017; 2018; Li et al., 2021; Huber et al., 2022), or (c) finding universal adversarial perturbations (Moosavi-Dezfooli et al.,

2017; Wallace et al., 2019; Mehrabi et al., 2022). Real attackers, meanwhile, are (i) known to make much larger edits from the original text, and (ii) are informed by each other's successful attacks, neither of which is captured in existing synthetic NLP attacks (West, 2017). In this paper, we take a step towards closing this gap by directly modeling the real attack patterns. This enables us to emulate human text attacks more realistically by (i) allowing larger edits and (ii) using existing real attacks to inform future attacks.

In prior work, the following proxies are usually used to measure whether generated adversarial examples are of good quality: semantic proximity (Malfa & Kwiatkowska, 2022), high attack success rate (Szegedy et al., 2014), low label noise (Malfa & Kwiatkowska, 2022), or distributional similarity to past attacks (Pillutla et al., 2021). These metrics, however, do not connect well to the attack patterns used by real adversaries. Our primary metric for attack quality is whether the generated attacks are useful when used in adversarial training to defend against future unseen rounds of human-generated attacks. That is, can we increase robustness beyond what we achieve by only training on all existing observed human attacks? We leverage frameworks like Dynabench to test on the evolving patterns of real adversaries (Kiela et al., 2021).

We show that our attack generation methods, which learn the distribution of human adversarial examples, outperform both (1) attack generators that do not take the human examples into account, and (2) attack generators that do not learn the distribution but rely on random perturbations (Sec 6). Our attack generation methods are able to make improvements even when trained on as few as 500 real human adversarial examples. Finally, though prior adversarial literature places a high emphasis on adversary success rates, low label noise, or distributional similarity, we show that these quality proxies are not predictive of whether an attack generator can better defend against future attacks. Our primary contributions are to:

1. **Demonstrate misalignment between synthetic and real attacks:** We empirically show that existing synthetic attack approaches do not necessarily improve robustness to the real attacks from humans.
2. **Overcome misalignment by imitating real adversaries:** We use generative models to directly imitate existing human-generated attacks. Our metric of success is how much we can improve robustness to future real attacks (beyond what can be accomplished by adversarially training on all existing real attacks).
3. **Improve adversarial robustness without relying on a better/bigger model:** Adversarial training on imitated real attacks provides significant robustness benefits. When compared to solely training on existing real attacks, we improve accuracy by 11% on unseen attacks in the ANLI benchmark, and by 8% on existing attacks in the hate speech detection benchmark.
4. **Show misalignment between common attack quality metrics and attack usefulness in preventing future attacks:** We empirically show that more distributional similarity, low label noise, or high adversary success rate do not entail that an attack generator is better than another in helping defend against downstream attacks.

2 Related Work

Adversarial robustness is measured as accuracy on challenge sets such as Adversarial GLUE (Wang et al., 2021), which are gathered through crowd-sourcing and programmatic text perturbation (Zhang et al., 2019; Morris et al., 2020b). To improve adversarial robustness, prior work uses training interventions and/or data augmentation. Training interventions focus on learning more meaningful stable representations from available data: e.g. using mutual information (Wang et al., 2020a; Zhao et al., 2022) or locally regularizing (Aghajanyan et al., 2020; Zhu et al., 2019). In contrast, we focus on data augmentation, which may be used with these training intervention methods.

Despite the popularity of data augmentation solutions such as ϵ-bounded PGD attacks (Madry et al., 2017) in continuous domains, it is not straightforward to extend them to discrete NLP settings. In NLP, small perturbations to the text can have a big impact on the text's true label.

Nevertheless, **controlled text generation** has vastly improved in recent years through zero or few-shot tuning of large language models (Wu et al., 2021; Perez et al., 2022). Such methods use data augmentation

for fine-tuning (Michel et al., 2019b; Garg & Ramakrishnan, 2020), gradients from small attribute classifiers to ensure generated text has a particular class (Dathathri et al., 2020; Wu et al., 2023), or careful prompting to generate semantically close text with only the desired attribute swapped (Madaan et al., 2020).

Adversarial controlled text generation, however, is is even more challenging. It's not straightforward to correctly label the generated text we intend to use as an adversarial example (by definition). Prior work typically gets around this challenge by making very small or rule-based perturbations that should not change the original example's true label. These include contextual synonym-substitution (Michel et al., 2019a; Jia et al., 2019b; Li et al., 2020a; Morris et al., 2020a), rule-based grammar (McCoy et al., 2019a; Ribeiro et al., 2018), morphological (Tan et al., 2020), and character-level (Eger & Benz, 2020) manipulations. A related but under-explored area is to adversarially intervene on all parts of the text that are invariant to the predicted label (Wang et al., 2020b; Chen et al., 2021; Lei et al., 2022), or conversely to minimally intervene to ensure the true label changes (Ross et al., 2021; Deng et al., 2022). Our attack method is different in that it relies on memorizing and re-mixing phrases from existing adversarial examples to generate additional adversarial examples that the attack method can correctly label.

Most of prior work assumes that adversarial text should be fluent, grammatically correct or semantically similar to the original. There is no such constraint on adversaries in the real world. Hence, in this paper we use the Hate Detection task which often does not have fluent grammatical text. Prior work that does not assume a standardized single language, include studies on dialectal language like Ebonics on Twitter (Blodgett et al., 2016), multilingual text on Reddit (Michel & Neubig, 2018), Hinglish (Biradar et al., 2021)), emoji-based hatespeech (Kirk et al., 2022); and methods that seek to distinguish human and machine generated text through heuristics (e.g.: sentence length, verb ratios, (Yao et al., 2017), relational consistency (Zhong et al., 2020)).

Red teaming: We are also motivated by red-team human generated datasets—human-created attacks that target a specific model, often by crowd-sourcing. Examples include SWAG, ReCoRD, HotpotQA, HellaSWAG, HANS (Zellers et al., 2018; Zhang et al., 2018; Yang et al., 2018; Zellers et al., 2019; McCoy et al., 2019a; Kaushik et al., 2019). Sometimes the work also uses feedback from the original classifier: CoDAH, Quoref, DROP, FEVER 2, ANLI, etc. (Chen et al., 2019; Dasigi et al., 2019; Dua et al., 2019; Thorne et al., 2019; Nie et al., 2020; Bartolo et al., 2020; Kiela et al., 2021). We rely on such human-model interactions in a feedback loop set-up, and propose a mechanism to reduce red team costs.

3 Problem Formulation

Training on existing real attacks improves robustness to future real attacks. Our goal is to improve robustness even further by making use of generated synthetic attacks that imitate the real adversarial examples. Given our focus on robustness improvements through data augmentation, we only assume black-box access to the attacked model, and knowledge of all observed existing attacks. We evaluate the robustness on successive rounds of adversarial examples collected through a crowd-sourced model-in-the-loop approach (Nie et al., 2020). We take the classifier architecture to be fixed; improvements to architecture or model size are out of scope.

Notation: We refer to $\mathbf{x}$, as a collection of samples drawn from distribution $p : \mathbf{x} \sim p$. We assume there are ground truth labels associated with samples $\mathbf{x}$, given by $y(\mathbf{x})$. We assume that there is a base classifier, parameterized by θ_0, trained on $\mathbf{x}$ that labels each input as $\widehat{y}_{\theta_0}(\mathbf{x})$. Next, we sample human adversarial examples from a distribution $p_{\mathbf{a}_0} : \mathbf{a} \sim p_{\mathbf{a}_0}$. These are created to fool the base classifier, i.e., $\widehat{y}_{\theta_0}(\mathbf{a}) \neq y(\mathbf{a})$.

The classifier is then further fine-tuned on the real attacks, $\mathbf{a}_0$. We refer to the output of this new classifier as $\widehat{y}_{\theta_1}(\mathbf{x})$. Next, we create a new round of human adversarial examples to fool $\widehat{y}_{\theta_1}$. We refer this future human adversarial data distribution as $p_{\mathbf{a}_1} : \mathbf{a}_1 \sim p_{\mathbf{a}_1}$ which are misclassified by the classifier: $\widehat{y}_{\theta_1}(\mathbf{a}_1) \neq y(\mathbf{a}_1)$.

This model-in-the-loop process results in evolving crowd-sourced real attacks as follows:

$$\theta_0 \rightarrow \mathbf{a}_0 \rightarrow \theta_1 \rightarrow \mathbf{a}_1 \rightarrow \cdots \tag{1}$$

We do not have access to $p_{\mathbf{a}_1}$,$p_{\mathbf{a}_2}$ or $\mathbf{a}_1$,$\mathbf{a}_2$ at train time. **Our goal is to improve the accuracy of $\widehat{y}_{\theta_1}$ on future attacks: $\mathbf{a}_1$, $\mathbf{a}_2$, without gathering additional human attacks beyond $\mathbf{a}_0$.**

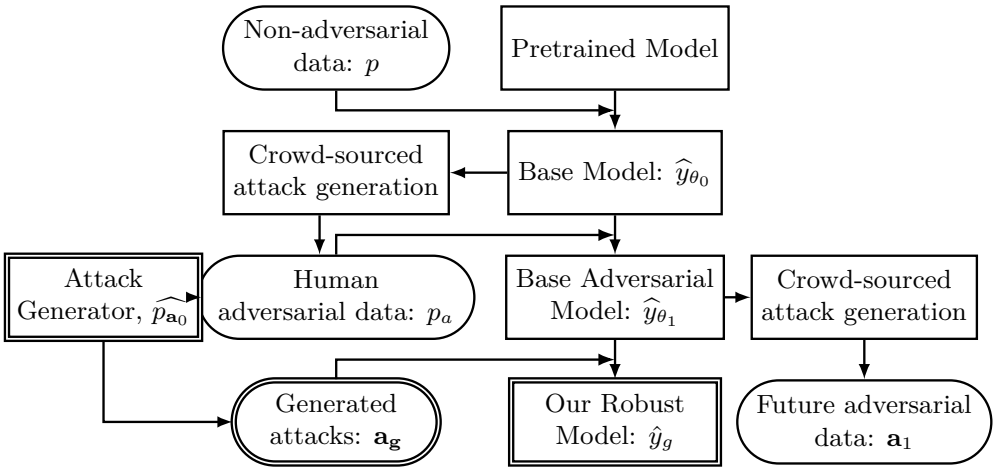

Figure 1: The framework that shows how each dataset and model is created. Our contributions are double lined. Not pictured: $\hat{y}_g$ is then evaluated on the future real attacks, $\mathbf{a}_1$

4 Methods

To address the problem in Sec. 3, we describe how we use synthetic attacks to improve robustness. Then we describe the specific methods of generating the synthetic attacks.

Overall Solution Framework

We aim to be more robust to $\mathbf{a}_1$ by fine-tuning $\widehat{y}_{\theta_1}$ on additional synthetic examples. We do this by training a generator, G, on existing adversarial examples, $\mathbf{a}_0 \sim p_{\mathbf{a}_0}$. The generator learns $\widehat{p_{\mathbf{a}_0}}$, an approximation of the true $p_{\mathbf{a}_0}$. Then we use the generator to create synthetic examples. $\mathbf{a_g} \sim \widehat{p_{\mathbf{a}_0}}$. (For clarity, we also depict how we create all models and adversarial data in Fig. 1.)

The hypothesis of this work is that modeling the existing real attack distribution allows the generator to capture something about the real attack generation process more broadly. This allows the generator to generalize to future attacks. **Restated, we hypothesize $\widehat{p_{\mathbf{a}_0}}$ is not only close to $p_{\mathbf{a}_0}$, but also reasonably close to future adversarial attack distributions, $p_{\mathbf{a}_1}$, $p_{\mathbf{a}_2}$.**

While in practice $\mathbf{a}_1$ and all subsequent attacks would depend on the mitigation put in place to obtain θ_1, we do not generate new human attacks in response to the new classifiers trained on $\mathbf{a_g}$, and hence $\mathbf{a}_1$ and $\mathbf{a}_2$ are fixed in our setting. We do show that future attacks ($\mathbf{a}_1$,$\mathbf{a}_2$) are less effective thanks to training on $\mathbf{a_g}$.

Synthetic Attack Generators

For many NLP tasks, the input x can be broken into (x_i, x_o). x_i is the portion of the input text that is not attacked by the adversary. It remains the same. This can be context: e.g. premise in NLI, paragraph in QA tasks, etc. For the toxicity task, where the entire sentence may be attacked, we set x_i to be the first half of the text. x_o is the portion that is attacked: e.g. hypothesis in NLI, question in QA, comment in sentiment analysis. We now present two methods to generate $\mathbf{a_g} \sim \widehat{p_{\mathbf{a}_0}}$. The first is a imitation-only approach agnostic to the task. The second takes the classifier task into account to better maintain the desired class label.

Direct Imitation (DI): Label-aware fine-tuning We fine-tune the generative model on existing observed attacks, $\mathbf{a}_0 \sim p_{\mathbf{a}_0}$. We consider (x_i, y) as the input for the generator, and x_o to be the target text that needs to be generated. Incorporating y as input greatly helps reduce the rate of noisy labels—the rate at which the the generated example does not retain the same label as the input ($y(\mathbf{a_g}) \neq y(\mathbf{a}_0)$). Specifically, we minimize the cross-entropy loss of the generated text probabilities $\widehat{p_{\mathbf{a}_0}}(x_i, y)$ and the target text x_o. See Appendix A for additional implementation details and loss function definition.

$\mathbf{a}_0$	The Nassau County population increased from 2010 to 2016.
$\mathbf{a}_0$	The Crystal Mountain Resort is a tourist destination.
$\mathbf{a_g}$	Crystal Mountain population increased from 2010 to 2016.

Table 1: An example of the ICE attack generator remixing existing observed attacks (top two) from the ANLI R1 data to create a new attack (bottom).

Imitation + Controlled Exploration (ICE): The primary challenge of the DI method, and the primary challenge of all controlled generation more broadly, is noisy labels (Also, adversarial examples are extra hard to label correctly by definition).

To overcome this challenge, we modify the Plug and Play controlled decoding method to make it suitable for adversarial robustness (Dathathri et al., 2020). Plug and Play methods add an additional small classifier to a generator. The classifier takes the hidden layers of the generator as input. This purpose of the classifier is to ensure a desired property is maintained in the generated output. For any output of the generator, the classifier checks whether the output has the desired property using a cross entropy loss. The generator updates its hidden layers to minimize the classifier loss, and generates a new better output that is closer to having the desired property. This method is computationally intensive if the classifier is complex. Hence, the classifier is a single layer feed forward network.

Warm-starting: This system works well when this classifier on the hidden layers of a generator model can capture the desired property; i.e the property is easy to learn. This is obviously not the case with adversarial examples. Restated, a linear classifier cannot guide the generator into producing correctly labeled examples when given tricky human adversarial examples as input. Moreover, it is unlikely that a linear classifier will outperform the large language model we aim to ultimately improve. Examples the classifier does label correctly, are examples the large language model is likely to get right as well. Hence, these examples will not be useful in improving the large language model.

We mitigate this issue by simplifying the task for the linear classifier by encouraging the generator to prefer text phrases for which the linear classifier already knows the true labels. Restated, the adversarial examples generated by this ICE method largely only re-mixes existing examples (See Table 1). For additional implementation details and loss function definition, we refer to the Appendix A.

We achieve this re-mixing by encouraging the system to reconstruct existing attacks. The reconstruction task asks the model to generate the attack exactly: $R : x_i \rightarrow x_o$. We modify the generator to multitask on observed real attacks, $\mathbf{a}_0$, for both the main task ($Y : x \rightarrow y$) and the reconstruction task. We freeze the generator parameters. Next, we fine-tune the linear classifier on the $\mathbf{a}_0$ for the main task. Then we freeze the parameters for the classifier and unfreeze the generator parameters.

Iterative Example Generation: We generate examples by passing in the observed real attacks one at a time. For a fixed number of S steps, the classifier provides feedback to the generator. The generator updates its parameters to generate a new output that is more likely to have the same label, y, as the input. This process is outlined in Algorithm 1: `LC` refers to the linear classifier; `G` refers to the generator. This method allows us to increase diversity of generated examples by toggling hyper-parameters: reducing the reconstruction task weight λ, increasing beam search parameter: α, and increasing steps: S.

5 Experiments

This section details how we evaluate the methods listed in Section 4—the benchmark dataset used, implementation details, and baselines.

A. Tasks: Human-in-the-loop crowd-sourced

We want to assess whether we can meaningfully amplify past human-generated attacks to be more robust to future human-generated attacks (given fixed attack generation instructions and UI). Hence, we chose two

Algorithm 1 Pseudo-code for ICE Method:

```
G.train(R : x_i → x_o, ∀(x, y) ∈ a_0)
G.train(Y : x → y, ∀(x, y) ∈ a_0)
Freeze weights of G.
Let H(x) = G.get_hidden_layers(R : x).
LC.train(Y : H(x) → y, ∀(x, y) ∈ a_0)
Freeze weights of LC
for all (x, y) ∈ a_0 do
    G' = G.copy()
    Unfreeze weights of G'
    for all i ∈ 1, 2, ··· , S do
        GRAD = ∇LC.loss(Y : (H(x_i), y)) + λ · ∇G.loss(R : (x_i, x_o))
        G'.back_propagate(GRAD)
    end for
    yield G'.decode(a_0, α)
end for
```

very different DynaBench tasks: Natural Language Inference (NLI) and Hate Speech Detection: The former has longer and qualitatively more varied texts. The latter is terse, less varied, and has less standard English (often with incorrect grammar and spelling) (Kiela et al., 2021).

A.I. Adversarial NLI: We evaluate our methods on the Adversarial NLI (ANLI) task (Nie et al., 2020). This is a Natural Language Inference (NLI) task: the goal is to determine whether a *hypothesis* logically follows (entailment) or contradicts (contradiction) or is undetermined (neutral) based on facts present in the *premise*. Nie et. al. crowd-source human attacks on the hypothesis against a base classifier, $\widehat{y}_{\theta_0}$, trained on MNLI+SNLI data (Wang et al., 2018; Bowman et al., 2015). Then they train more robust models by incorporating these new attacks (and other data) for three rounds. $\mathbf{a}_0$ are the human generated attacks from the first round created by attacking $\widehat{y}_{\theta_0}$, a BERT-Large transformer classifier (Devlin et al., 2019). Successive rounds $\mathbf{a}_1$, $\mathbf{a}_2$ are created by attacking RoBERTa models. We choose to improve the robustness of BERT-Large model $\widehat{y}_{\theta_0}$ using $\mathbf{a}_0$ and evaluate on future human adversarial attacks: $\mathbf{a}_1$,$\mathbf{a}_2$.

A.II. Hate Speech Detection: We also evaluate on the Dynabench Hate Speech detection dataset, an adversarial human-in-the-loop dataset generated in four rounds (Vidgen et al., 2021). In Round 1, Vidgen et. al. train a base RoBERTA classifier, $\widehat{y}_{\theta_0}$, on original content created by humans. In Rounds 2-4, they create more robust RoBERTa models ($\widehat{y}_{\theta_1}$,$\widehat{y}_{\theta_2}$,$\widehat{y}_{\theta_3}$) by training on attacks created as follows: Human raters first create original content that successfully fools the base classifier. Then they perturb these new sentences to create even more challenging "contrast sets" with different labels. This data is then split into train, validation, and test sets, with half the test set entries created by annotators who do not appear in the training and validation sets to minimize annotator bias.

Note that for both tasks we do not gather additional human adversarial attacks targeting our improved classifiers. We evaluate on a fixed set of attacks previously unseen by the classifier. This is a limitation of our set-up; gathering additional rounds of human attacks is left as future work.

B. Base Attack Generator Model is T5

We use the T5 encoder-decoder as our generator for this paper (Raffel et al., 2020b) for two reasons. First, it is compatible with multiple small sentence generation tasks (Raffel et al., 2020a). Second, its performance on benchmark NLI tasks (MNLI, ANLI), and the Hate Speech task is close to that of the BERT-Large model we seek to improve – i.e. improvements are not coming from a superior model but rather imitating the real adversaries.

Model	$\mathbf{a}_0$: R1	$\mathbf{a}_1$: R2	$\mathbf{a}_2$: R3
$\widehat{y}_{\theta_1}$: Base + R1	$44.1_{\pm 0.03}$	$32.5_{\pm 0.05}$	$29.4_{\pm 0.05}$
↪ + TextFooler(R1)	$24.1_{\pm 0.08}$	$27.9_{\pm 0.06}$	$30.3_{\pm 0.06}$
↪ + BERT-Attack(R1)	$35.1_{\pm 0.12}$	$29.0_{\pm 0.08}$	$31.3_{\pm 0.09}$
↪ + CT-GAN(R1)	$26.8_{\pm 0.14}$	$29.5_{\pm 0.12}$	$29.5_{\pm 0.11}$
↪ + DI(MNLI+SNLI)	$22.9_{\pm 0.11}$	$28.1_{\pm 0.12}$	$29.4_{\pm 0.10}$
↪ + ICE(MNLI+SNLI)	$33.9_{\pm 0.78}$	$\underline{33.7}_{\pm 0.67}$	$\underline{33.5}_{\pm 1.47}$
↪ + DI(R1)	$\underline{48.2}_{\pm 0.32}$	$\underline{39.1}_{\pm 0.29}$	$\underline{\mathbf{40.2}}_{\pm 0.37}$
↪ + ICE(R1)	$\underline{\mathbf{50.1}}_{\pm 1.43}$	$\underline{\mathbf{43.4}}_{\pm 2.91}$	$\underline{\mathbf{39.9}}_{\pm 1.38}$

Table 2: Improvement on ANLI mean accuracy (%) (± standard error across 3 runs) when trained on attacks generated only from Round 1. The notation, DI(R1) for instance, refers to the method DI using R1 data to generate more examples. We underline the setups that outperform the *Base + R1* baseline.

C. Baselines

We have two types of baseline attack generators: The first type uses existing observed attacks, but does not learn the distribution, instead relying on random perturbations. We use three of these methods: TextFooler, BertAttack, and CT-GAN. TextFooler is a very popular attack generation library that transmutes the most predictive words, while preserving semantic similarity and contextual coherence (Jin et al., 2020). BertAttack is another popular method: it uses the model it's attacking to identify vulnerable words in the input; then, it uses BERT to generate substitutes for the vulnerable words. (Li et al., 2020b). CT-GAN is a Generative Adversarial Network(Goodfellow et al., 2014) modified for controlled text generation where the NLI premise is used as the control text (Haidar et al., 2019; Betti et al., 2020). The second type of baseline learns an example distribution, but does not use the attack distribution. We repeat our main methods, ICE and DI, in a new data setting. Instead of using ANLI R1, we use the MNLI+SNLI data used in to train the base classifier, $\widehat{y}_{\theta_0}$.

6 Results

A. Synthetic human-like adversarial data improve robustness to future attacks. Distribution-agnostic baselines do not.

Table 2 and Table 3 show that for both tasks, the accuracy on future rounds of human-generated attack ($\mathbf{a}_1$, $\mathbf{a}_2$) improves when generated examples, $\mathbf{a_g}$ from ICE and DI are incorporated into training.

The adversarial example generators that attempt to imitate $\mathbf{a}_1$ (ICE and DI) out-perform all types of baselines. First, they improve robustness beyond what we achieve by training on past human adversarial attacks, $\mathbf{a}_1$, alone. This improvement cannot be achieved merely by training for more steps on ANLI R1 as shown in Table 18 in the Appendix. Second, they out-perform methods that rely on noise-based attacks on $\mathbf{a}_1$ like TextFooler, BertAttack, and CT-GAN. Finally, they out-perform methods that imitate example distributions generated by other processes: ICE(MNLI+SNLI) and DI(MNLI+SNLI). The word/phrase substitution methods, BertAttack and TextFooler, improve accuracy within the same round for the Hate Speech dataset–this dataset is itself half-generated by making such minimal substitutions. Yet, these methods are not more effective on future rounds.

Table 17 extends the setting where R1 and R2 are past observed attacks available for training, and R3 is the held out set of future human attacks, leading to similar conclusions.

B. Common metrics for an attack method (label noise rate, attack success rate, and distributional similarity) do not entail whether adversarial data from the method can help defend against future attacks.

Adversarial robustness literature considers attack methods to be better if they produce datasets with less label noise (Dathathri et al., 2020), or higher attack success rates (Uesato et al., 2018), or higher proximity to the original dataset (Ross et al., 2021). We show that these metrics are not good proxies for determining which type of attack method can generate examples that best defends against future attacks. These findings

Model	a_0: R2	a_1: R3	a_2: R4
Base + R1 + R2	$0.76_{\pm 0.001}$	$0.78_{\pm 0.003}$	$0.77_{\pm 0.001}$
↪ + TextFooler(R2)	$\underline{0.78}_{\pm 0.011}$	$0.77_{\pm 0.012}$	$0.76_{\pm 0.009}$
↪ + BERT-Attack(R2)	$\underline{0.78}_{\pm 0.013}$	$0.76_{\pm 0.015}$	$0.77_{\pm 0.017}$
↪ + DI(R2)	$\underline{\mathbf{0.84}}_{\pm 0.013}$	$0.77_{\pm 0.034}$	$0.76_{\pm 0.018}$
↪ + ICE(R2)	$\underline{\mathbf{0.83}}_{\pm 0.032}$	$0.80_{\pm 0.024}$	$\underline{\mathbf{0.79}}_{\pm 0.018}$

Table 3: Improvement on Hate speech detection AUC (± standard error across 3 runs) when trained on attacks generated only from Round 2. The notation, DI(R2) for instance, refers to the method DI using R2 data to generate more examples. In this dataset, R1 is not adversarially generated, and is analogous to the base MNLI/SNLI data in the ANLI task.

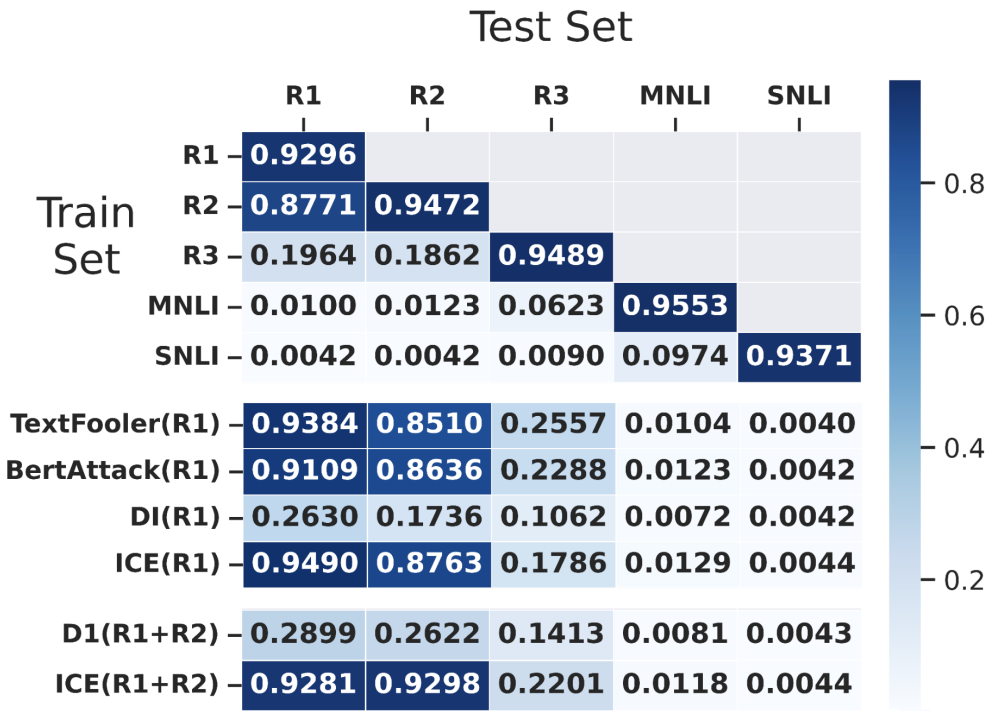

Figure 2: Distributional similarity, as measured by MAUVE on RoBERTa embeddings from a random 1k sample (Pillutla et al., 2021). MAUVE scores range from 0 to 1, with higher values indicating more similar distributions. MAUVE metrics are intended to be evaluated relative to each other, and not as absolute measures. Note that distributional similarity to the held out attacks, R2, does not correlate with whether an attack generation method is useful as per Table 2.

are surprising, and additional investigation is needed to understand why (Also see generated examples in Appendix A).

B.I. Higher Distributional Similarity of a_g to future attacks does not entail more useful adversarial examples. We use MAUVE as a metric of distributional similarity between two text datasets in Fig. 2 (Pillutla et al., 2021). Though DI(R1) and ICE(R1) are more useful methods than TextFooler(R1) (as per Table 2), ICE and TextFooler have about the same level of similarity to the ANLI data — both higher than DI.

B.II. Lower label noise does not entail more useful adversarial examples. One of the hardest challenges in controlled adversarial example generation is generating examples that have the desired label, that is reducing the rate of noisy labels. While this is a useful goal within an attack method type, comparing rates of noisy labels across attack methods does not help us choose a more useful method. Table 4 shows

Generated attack dataset	Rate of correct labels
$\mathbf{a_g}$: TextFooler(R1)	51%
$\mathbf{a_g}$: DI(R1)	39%
$\mathbf{a_g}$: ICE(R1)	**78%**

Table 4: Rate of correct labels from the attack generation methods on ANLI round 1 (↑ better). We, the authors, rated 100 random generated examples from each method to obtain these numbers. Our rating guidelines are included in the Appendix A, and the ratings are in Supplemental Materials.

Model to attack → **Attack Dataset ↓**	Base $\widehat{y}_{\theta_0}$	Base+R1 $\widehat{y}_{\theta_1}$	Base+R1+R2 $\widehat{y}_{\theta_2}$
$\mathbf{a}_0$: R1	78%	56%	47%
$\mathbf{a}_1$: R2	73%	67%	58%
$\mathbf{a}_2$: R3	71%	71%	62%
$\mathbf{a_g}$: TextFooler(R1)	96%	98%	94%
$\mathbf{a_g}$: DI(R1)	76%	87%	79%
$\mathbf{a_g}$: ICE(R1)	78%	88%	88%

Table 5: Attack success rate of attack datasets on base BERT Large models as they incorporate more rounds of ANLI training data (↑ better). Only the examples that had the correct labels (as verified by human rating) are included in computing this rate.

that TextFooler has more correct labels than the DI method, even though Table 2 makes it clear that DI is the more useful method. We, the authors, annotated 100 examples from each attack method and report the rate of correct labels by comparing our annotated ground truth with the adversarial label associated with the corresponding (premise, hypothesis) pair in the ANLI R1 dataset. We refer to the Appendix A for tables of generated examples from each of the attack methods, and guidelines for human annotations.

B.III. Higher adversary success rate does not entail more useful adversarial examples. Table 5 shows the attack success rate of the various attack datasets (only including attacks with the correct label as verified by human ratings). All synthetic attack generators have high success rates all classifiers - which indicates that the information in the generated examples were not yet captured in by the base classifiers. Surprisingly, TextFooler, which has the highest attack success rate, does not improve adversarial robustness; R2, R3, DI(R1), and ICE(R1) datasets have lower attack success rates but are much more useful in increasing robustness to future attacks.

C. Even ∼1k human adversarial examples improves robustness to unseen adversaries.

We test how the number of human adversarial examples used to train the attack generator affects the adversarial robustness using the generated examples. We fix the number of examples we use to train $\widehat{y}_{\theta_1}$ at 10k generated examples. The full set of real adversarial examples from ANLI R1 is 16.9k real examples. We sample the number of real examples we use to train the attack generator. Table 6 demonstrates that increasing human examples improves robustness. Nevertheless, even when we provide low numbers of human examples, robustness to future rounds has improved. In-distribution accuracy on R1, however, suffers until 8k examples are used to train the attack generator. When we have fewer than 500 human adversarial examples to amplify using our approach, the robustness gains do not generalize. That happens because the few-shot setting creates several additional considerations that are out-of-scope in this paper:

- Simple fine-tuning baselines do not perform well in the few-shot setting. Other few-shot and parameter efficient baselines might be relevant (Zhou et al., 2022; Liu et al., 2022). The choice of training method is orthogonal to our goal of synthetic data generation.

# R1 samples	a_0: R1	a_2: R2	a_2: R3
100	33.0	27.8	28.9
200	34.2	28.1	29.5
500	37.1	30.8	32.6
1024	42.7	36.5	36.1
2048	43.2	33.2	36.5
4096	43.3	37.8	37.1
8192	44.5	36.6	36.6
all (16.9k)	**48.2**	**39.1**	**40.1**

Table 6: Test accuracy (%) as we vary the number of R1 real adversarial examples used to fine-tune the DI attack generator. We use the trained DI generator to generate 10k examples. We fine-tune a Base + R1 classifier on these 10k DI(R1) examples to produce the robustness metrics above. We underline the setups that outperform the *Base + R1* baseline in Table 2.

- As we reduce the number of adversarial examples, evaluating generalization requires categorizing examples into various patterns to ensure coverage. In this work, we instead use the pattern-agnostic approach of Dynabench (Kiela et al., 2021), which aggregates all human-generated attacks.

.

D. The ICE method, in effect, re-mixes phrases from previous observed attacks.

The reconstruction loss (R) often leads to exact memorization of train-set examples or exact phrases from the train-set. Qualitatively, we find that increasing the beam search parameter α and number of steps S, or decreasing reconstruction loss, λ at inference time leads to remixing these phrases (See Table 1). This also explains the high MAUVE distributional similarity between the ICE methods and the ANLI rounds on which it trains.

7 Limitations

If we put our new more robust models to use, human adversaries may adapt to them as well. Checking whether crowd-sourcing fresh attacks is indeed more difficult on the new models is beyond the scope of this work. Also, we benefit from having a fair number of human adversarial examples (16.9k in ANLI, and 10k per round in Hate Speech). Our methods may be less successful in a scenario with very fewer examples ($\sim$ 10). On the flip side, we have also not evaluated these methods on classifiers with access to even larger real adversarial datasets. Finally, our methods work on datasets with the notion of an original example, and a perturbed adversarial example as is the norm for adversarial robustness literature (Madry et al., 2017). In the new paradigm of larger more capable NLP models, adversarial datasets may increasingly not involve a perturbation (Ganguli et al., 2022b).

Risks: Our techniques can enhance robustness given a set of observed adversarial examples. The new classifier we trained with generated data from DI and ICE may still be vulnerayble to future human attacks that are able to adapt to the new model (in this paper future attack rounds are known a priori from past model-in-the-loop work). This would require extensive crowd-sourcing efforts to evaluate. We also run the risk of over-fitting to the new human generated adversarial data. This may come at the cost of lower performance on future attacks generated by a different mechanism (say TextFooler instead of future ANLI rounds), and comes at the cost of degrading accuracy on original tasks such as MNLI and SNLI (Table 19 in Appendix). As a separate concern, any technique that betters generative text modeling brings the risk that humans may struggle to distinguish machine generated text. This can have negative consequences for disinformation and misinformation, which is an active area of research (Pu et al., 2023).

8 Conclusion

We demonstrate that training on attacks that imitate human adversaries can improve robustness to future rounds of human adversarial attacks by 11% on ANLI, and by 6% and 8% on existing adversarial examples

on ANLI and hate speech datasets. We are able to improve robustness to future attack distributions even when the attack generator is only trained on 1000 real adversarial examples. We show that existing attack generation methods that do not train on the distribution of real attacks, however, (methods like TextFooler and CT-GANs) are unable to improve robustness to future real attacks. Finally, we discover that attack generation methods with the lowest label noise or highest attack success rate or highest distributional similarity to future attacks are not the best methods at increasing robustness to future real attacks. These findings run counter to accepted norms on choosing the best attack method type in literature, demonstrate the opportunities in leveraging increasingly effective generation methods, and motivate future work on improving real adversarial robustness.

References

Armen Aghajanyan, Akshat Shrivastava, Anchit Gupta, Naman Goyal, Luke Zettlemoyer, and Sonal Gupta. Better fine-tuning by reducing representational collapse. *arXiv preprint arXiv:2008.03156*, 2020.

Max Bartolo, Alastair Roberts, Johannes Welbl, Sebastian Riedel, and Pontus Stenetorp. Beat the AI: Investigating adversarial human annotation for reading comprehension. *Transactions of the Association for Computational Linguistics*, 8:662–678, 2020. doi: 10.1162/tacl_a_00338. URL https://aclanthology.org/2020.tacl-1.43.

Federico Betti, Giorgia Ramponi, and Massimo Piccardi. Controlled text generation with adversarial learning. In *Proceedings of the 13th International Conference on Natural Language Generation*, pp. 29–34, Dublin, Ireland, December 2020. Association for Computational Linguistics. URL https://aclanthology.org/2020.inlg-1.5.

Shankar Biradar, Sunil Saumya, and Arun Chauhan. Hate or non-hate: Translation based hate speech identification in code-mixed hinglish data set. In *2021 IEEE International Conference on Big Data (Big Data)*, pp. 2470–2475, 2021. doi: 10.1109/BigData52589.2021.9671526.

Su Lin Blodgett, Lisa Green, and Brendan O'Connor. Demographic dialectal variation in social media: A case study of African-American English. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 1119–1130, Austin, Texas, November 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1120. URL https://aclanthology.org/D16-1120.

Daniel Borkan, Lucas Dixon, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. Nuanced metrics for measuring unintended bias with real data for text classification. In *Companion proceedings of the 2019 world wide web conference*, pp. 491–500, 2019.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 632–642, Lisbon, Portugal, September 2015. Association for Computational Linguistics. doi: 10.18653/v1/D15-1075. URL https://aclanthology.org/D15-1075.

Michael Chen, Mike D'Arcy, Alisa Liu, Jared Fernandez, and Doug Downey. CODAH: An adversarially-authored question answering dataset for common sense. In *Proceedings of the 3rd Workshop on Evaluating Vector Space Representations for NLP*, pp. 63–69, Minneapolis, USA, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-2008. URL https://aclanthology.org/W19-2008.

Yangyi Chen, Jin Su, and Wei Wei. Multi-granularity textual adversarial attack with behavior cloning. *arXiv preprint arXiv:2109.04367*, 2021.

Pradeep Dasigi, Nelson F. Liu, Ana Marasović, Noah A. Smith, and Matt Gardner. Quoref: A reading comprehension dataset with questions requiring coreferential reasoning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 5925–5932, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1606. URL https://aclanthology.org/D19-1606.

Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. Plug and play language models: A simple approach to controlled text generation. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=H1edEyBKDS.

Chuyun Deng, Mingxuan Liu, Yue Qin, Jia Zhang, Hai-Xin Duan, and Donghong Sun. ValCAT: Variable-length contextualized adversarial transformations using encoder-decoder language model. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 1735–1746, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.125. URL https://aclanthology.org/2022.naacl-main.125.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423.

Emily Dinan, Samuel Humeau, Bharath Chintagunta, and Jason Weston. Build it break it fix it for dialogue safety: Robustness from adversarial human attack. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 4537–4546, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1461. URL https://www.aclweb.org/anthology/D19-1461.

Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. DROP: A reading comprehension benchmark requiring discrete reasoning over paragraphs. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 2368–2378, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1246. URL https://aclanthology.org/N19-1246.

Steffen Eger and Yannik Benz. From hero to zéroe: A benchmark of low-level adversarial attacks. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pp. 786–803, Suzhou, China, December 2020. Association for Computational Linguistics. URL https://aclanthology.org/2020.aacl-main.79.

Allyson Ettinger, Sudha Rao, Hal Daumé III, and Emily M. Bender. Towards linguistically generalizable NLP systems: A workshop and shared task. In *Proceedings of the First Workshop on Building Linguistically Generalizable NLP Systems*, pp. 1–10, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-5401. URL https://aclanthology.org/W17-5401.

Deep Ganguli, Liane Lovitt, Jackson Kernion, Amanda Askell, Yuntao Bai, Saurav Kadavath, Ben Mann, Ethan Perez, Nicholas Schiefer, Kamal Ndousse, Andy Jones, Sam Bowman, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Nelson Elhage, Sheer El-Showk, Stanislav Fort, Zac Hatfield Dodds, Tom Henighan, Danny Hernandez, Tristan Hume, Josh Jacobson, Scott Johnston, Shauna Kravec, Catherine Olsson, Sam Ringer, Eli Tran-Johnson, Dario Amodei, Tom Brown, Nicholas Joseph, Sam McCandlish, Chris Olah, Jared Kaplan, and Jack Clark. Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned, 2022a. URL https://arxiv.org/abs/2209.07858.

Deep Ganguli, Liane Lovitt, Jackson Kernion, Amanda Askell, Yuntao Bai, Saurav Kadavath, Ben Mann, Ethan Perez, Nicholas Schiefer, Kamal Ndousse, Andy Jones, Sam Bowman, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Nelson Elhage, Sheer El-Showk, Stanislav Fort, Zac Hatfield-Dodds, Tom Henighan, Danny Hernandez, Tristan Hume, Josh Jacobson, Scott Johnston, Shauna Kravec, Catherine Olsson, Sam Ringer, Eli Tran-Johnson, Dario Amodei, Tom Brown, Nicholas Joseph, Sam McCandlish, Chris Olah, Jared Kaplan, and Jack Clark. Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned, 2022b.

Ji Gao, Jack Lanchantin, Mary Lou Soffa, and Yanjun Qi. Black-box generation of adversarial text sequences to evade deep learning classifiers. *CoRR*, abs/1801.04354, 2018. URL http://arxiv.org/abs/1801.04354.

Siddhant Garg and Goutham Ramakrishnan. Bae: Bert-based adversarial examples for text classification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 6174–6181, 2020.

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks, 2014. URL https://arxiv.org/abs/1406.2661.

Md. Akmal Haidar, Mehdi Rezagholizadeh, Alan Do Omri, and Ahmad Rashid. Latent code and text-based generative adversarial networks for soft-text generation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 2248–2258, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1234. URL https://aclanthology.org/N19-1234.

Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15262–15271, June 2021.

Lukas Huber, Marc Alexander Kühn, Edoardo Mosca, and Georg Groh. Detecting word-level adversarial text attacks via SHapley additive exPlanations. In *Proceedings of the 7th Workshop on Representation Learning for NLP*, pp. 156–166, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.repl4nlp-1.16. URL https://aclanthology.org/2022.repl4nlp-1.16.

Robin Jia and Percy Liang. Adversarial examples for evaluating reading comprehension systems, 2017. URL https://arxiv.org/abs/1707.07328.

Robin Jia, Aditi Raghunathan, Kerem Göksel, and Percy Liang. Certified robustness to adversarial word substitutions. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 4129–4142, Hong Kong, China, November 2019a. Association for Computational Linguistics. doi: 10.18653/v1/D19-1423. URL https://aclanthology.org/D19-1423.

Robin Jia, Aditi Raghunathan, Kerem Göksel, and Percy Liang. Certified robustness to adversarial word substitutions, 2019b.

Di Jin, Zhijing Jin, Joey Tianyi Zhou, and Peter Szolovits. 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*, volume 34, pp. 8018–8025, 2020.

Divyansh Kaushik, Eduard Hovy, and Zachary C. Lipton. Learning the difference that makes a difference with counterfactually-augmented data, 2019. URL https://arxiv.org/abs/1909.12434.

Douwe Kiela, Max Bartolo, Yixin Nie, Divyansh Kaushik, Atticus Geiger, Zhengxuan Wu, Bertie Vidgen, Grusha Prasad, Amanpreet Singh, Pratik Ringshia, Zhiyi Ma, Tristan Thrush, Sebastian Riedel, Zeerak Waseem, Pontus Stenetorp, Robin Jia, Mohit Bansal, Christopher Potts, and Adina Williams. Dynabench: Rethinking benchmarking in NLP. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4110–4124, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.324. URL https://aclanthology.org/2021.naacl-main.324.

Hannah Kirk, Bertie Vidgen, Paul Rottger, Tristan Thrush, and Scott Hale. Hatemoji: A test suite and adversarially-generated dataset for benchmarking and detecting emoji-based hate. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 1352–1368, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.97. URL https://aclanthology.org/2022.naacl-main.97.

Alyssa Lees, Daniel Borkan, Ian Kivlichan, Jorge Nario, and Tesh Goyal. Capturing covertly toxic speech via crowdsourcing. In *Proceedings of the First Workshop on Bridging Human–Computer Interaction and Natural Language Processing*, pp. 14–20, 2021.

Yibin Lei, Yu Cao, Dianqi Li, Tianyi Zhou, Meng Fang, and Mykola Pechenizkiy. Phrase-level textual adversarial attack with label preservation. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pp. 1095–1112, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-naacl.83. URL https://aclanthology.org/2022.findings-naacl.83.

Dianqi Li, Yizhe Zhang, Hao Peng, Liqun Chen, Chris Brockett, Ming-Ting Sun, and Bill Dolan. Contextualized perturbation for textual adversarial attack. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 5053–5069, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.400. URL https://aclanthology.org/2021.naacl-main.400.

Linyang Li, Ruotian Ma, Qipeng Guo, X. Xue, and Xipeng Qiu. Bert-attack: Adversarial attack against bert using bert. *ArXiv*, abs/2004.09984, 2020a.

Linyang Li, Ruotian Ma, Qipeng Guo, Xiangyang Xue, and Xipeng Qiu. BERT-ATTACK: Adversarial attack against BERT using BERT. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 6193–6202, Online, November 2020b. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.500. URL https://aclanthology.org/2020.emnlp-main.500.

Haokun Liu, Derek Tam, Mohammed Muqeeth, Jay Mohta, Tenghao Huang, Mohit Bansal, and Colin A Raffel. Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning. *Advances in Neural Information Processing Systems*, 35:1950–1965, 2022.

Nishtha Madaan, Inkit Padhi, Naveen Panwar, and Diptikalyan Saha. Generate your counterfactuals: Towards controlled counterfactual generation for text. *CoRR*, abs/2012.04698, 2020. URL https://arxiv.org/abs/2012.04698.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2017.

Emanuele La Malfa and Marta Z. Kwiatkowska. The king is naked: on the notion of robustness for natural language processing. In *AAAI*, 2022.

Tom McCoy, Ellie Pavlick, and Tal Linzen. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 3428–3448, Florence, Italy, July 2019a. Association for Computational Linguistics. doi: 10.18653/v1/P19-1334. URL https://aclanthology.org/P19-1334.

Tom McCoy, Ellie Pavlick, and Tal Linzen. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 3428–3448, 2019b.

Ninareh Mehrabi, Ahmad Beirami, Fred Morstatter, and Aram Galstyan. Robust conversational agents against imperceptible toxicity triggers. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 2831–2847, 2022. URL https://aclanthology.org/2022.naacl-main.204/.

Paul Michel and Graham Neubig. MTNT: A testbed for machine translation of noisy text. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 543–553, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1050. URL https://aclanthology.org/D18-1050.

Paul Michel, Xian Li, Graham Neubig, and Juan Pino. On evaluation of adversarial perturbations for sequence-to-sequence models. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 3103–3114, Minneapolis, Minnesota, June 2019a. Association for Computational Linguistics. doi: 10.18653/v1/N19-1314. URL https://aclanthology.org/N19-1314.

Paul Michel, Xian Li, Graham Neubig, and Juan Miguel Pino. On evaluation of adversarial perturbations for sequence-to-sequence models, 2019b. URL https://arxiv.org/abs/1903.06620.

Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Omar Fawzi, and Pascal Frossard. Universal adversarial perturbations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1765–1773, 2017.

John Morris, Eli Lifland, Jack Lanchantin, Yangfeng Ji, and Yanjun Qi. Reevaluating adversarial examples in natural language. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 3829–3839, Online, November 2020a. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.341. URL https://aclanthology.org/2020.findings-emnlp.341.

John Morris, Eli Lifland, Jin Yong Yoo, Jake Grigsby, Di Jin, and Yanjun Qi. Textattack: A framework for adversarial attacks, data augmentation, and adversarial training in nlp. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 119–126, 2020b.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. Adversarial NLI: A new benchmark for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4885–4901, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.441. URL https://aclanthology.org/2020.acl-main.441.

Ethan Perez, Saffron Huang, Francis Song, Trevor Cai, Roman Ring, John Aslanides, Amelia Glaese, Nat McAleese, and Geoffrey Irving. Red teaming language models with language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 3419–3448, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.emnlp-main.225.

Krishna Pillutla, Swabha Swayamdipta, Rowan Zellers, John Thickstun, Sean Welleck, Yejin Choi, and Zaid Harchaoui. Mauve: Measuring the gap between neural text and human text using divergence frontiers. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 4816–4828. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper/2021/file/260c2432a0eecc28ce03c10dadc078a4-Paper.pdf.

J. Pu, Z. Sarwar, S. Muhammad Abdullah, A. Rehman, Y. Kim, P. Bhattacharya, M. Javed, and B. Viswanath. Deepfake text detection: Limitations and opportunities. In *2023 2023 IEEE Symposium on Security and Privacy (SP) (SP)*, pp. 19–36, Los Alamitos, CA, USA, may 2023. IEEE Computer Society. doi: 10.1109/SP46215.2023.00002. URL https://doi.ieeecomputersociety.org/10.1109/SP46215.2023.00002.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020a. URL http://jmlr.org/papers/v21/20-074.html.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020b. URL http://jmlr.org/papers/v21/20-074.html.

Shuhuai Ren, Yihe Deng, Kun He, and Wanxiang Che. Generating natural language adversarial examples through probability weighted word saliency. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 1085–1097, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1103. URL https://aclanthology.org/P19-1103.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Semantically equivalent adversarial rules for debugging NLP models. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 856–865, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1079. URL https://aclanthology.org/P18-1079.

Alexis Ross, Ana Marasović, and Matthew Peters. Explaining NLP models via minimal contrastive editing (MiCE). In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pp. 3840–3852, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-acl.336. URL https://aclanthology.org/2021.findings-acl.336.

Morgan Klaus Scheuerman, Jialun Aaron Jiang, Casey Fiesler, and Jed R. Brubaker. A framework of severity for harmful content online. *Proc. ACM Hum.-Comput. Interact.*, 5(CSCW2), oct 2021. doi: 10.1145/3479512. URL https://doi.org/10.1145/3479512.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *International Conference on Learning Representations*, 2014. URL http://arxiv.org/abs/1312.6199.

Samson Tan, Shafiq Joty, Min-Yen Kan, and Richard Socher. It's morphin' time! Combating linguistic discrimination with inflectional perturbations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 2920–2935, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.263. URL https://aclanthology.org/2020.acl-main.263.

James Thorne, Andreas Vlachos, Oana Cocarascu, Christos Christodoulopoulos, and Arpit Mittal. The FEVER2.0 shared task. In *Proceedings of the Second Workshop on Fact Extraction and VERification (FEVER)*, pp. 1–6, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-6601. URL https://aclanthology.org/D19-6601.

Jonathan Uesato, Brendan O'Donoghue, Aaron van den Oord, and Pushmeet Kohli. Adversarial risk and the dangers of evaluating against weak attacks, 2018.

Bertie Vidgen, Tristan Thrush, Zeerak Waseem, and Douwe Kiela. Learning from the worst: Dynamically generated datasets to improve online hate detection, 2021.

Eric Wallace, Shi Feng, Nikhil Kandpal, Matt Gardner, and Sameer Singh. Universal adversarial triggers for attacking and analyzing nlp. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2153–2162, 2019.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. *CoRR*, abs/1804.07461, 2018. URL http://arxiv.org/abs/1804.07461.

Boxin Wang, Shuohang Wang, Yu Cheng, Zhe Gan, Ruoxi Jia, Bo Li, and Jingjing Liu. Infobert: Improving robustness of language models from an information theoretic perspective. *arXiv preprint arXiv:2010.02329*, 2020a.

Boxin Wang, Chejian Xu, Shuohang Wang, Zhe Gan, Yu Cheng, Jianfeng Gao, Ahmed Hassan Awadallah, and Bo Li. Adversarial glue: A multi-task benchmark for robustness evaluation of language models. *arXiv preprint arXiv:2111.02840*, 2021.

Tianlu Wang, Xuezhi Wang, Yao Qin, Ben Packer, Kang Li, Jilin Chen, Alex Beutel, and Ed H. Chi. Cat-gen: Improving robustness in nlp models via controlled adversarial text generation. In *EMNLP*, 2020b.

Sarah West. Raging against the machine: Network gatekeeping and collective action on social media platforms. *Media and Communication*, 5(3):28–36, 2017. ISSN 2183-2439. doi: 10.17645/mac.v5i3.989. URL https://www.cogitatiopress.com/mediaandcommunication/article/view/989.

Hongqiu Wu, Yongxiang Liu, Hanwen Shi, Hai Zhao, and Min Zhang. Toward adversarial training on contextualized language representation. *arXiv preprint arXiv:2305.04557*, 2023.

Tongshuang Wu, Marco Tulio Ribeiro, Jeffrey Heer, and Daniel Weld. Polyjuice: Generating counterfactuals for explaining, evaluating, and improving models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 6707–6723, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.523. URL https://aclanthology.org/2021.acl-long.523.

Jing Xu, Da Ju, Margaret Li, Y-Lan Boureau, Jason Weston, and Emily Dinan. Bot-adversarial dialogue for safe conversational agents. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 2950–2968, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.235. URL https://aclanthology.org/2021.naacl-main.235.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. *CoRR*, abs/1809.09600, 2018. URL http://arxiv.org/abs/1809.09600.

Yuanshun Yao, Bimal Viswanath, Jenna Cryan, Haitao Zheng, and Ben Y Zhao. Automated crowdturfing attacks and defenses in online review systems. In *Proceedings of the 2017 ACM SIGSAC conference on computer and communications security*, pp. 1143–1158, 2017.

Yuan Zang, Fanchao Qi, Chenghao Yang, Zhiyuan Liu, Meng Zhang, Qun Liu, and Maosong Sun. Word-level textual adversarial attacking as combinatorial optimization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 6066–6080, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.540. URL https://aclanthology.org/2020.acl-main.540.

Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. SWAG: A large-scale adversarial dataset for grounded commonsense inference. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 93–104, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1009. URL https://aclanthology.org/D18-1009.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. HellaSwag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4791–4800, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1472. URL https://aclanthology.org/P19-1472.

Sheng Zhang, Xiaodong Liu, Jingjing Liu, Jianfeng Gao, Kevin Duh, and Benjamin Van Durme. Record: Bridging the gap between human and machine commonsense reading comprehension. *CoRR*, abs/1810.12885, 2018. URL http://arxiv.org/abs/1810.12885.

Wei Emma Zhang, Quan Z Sheng, Ahoud Alhazmi, and Chenliang Li. Adversarial attacks on deep-learning models in natural language processing: A survey. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 11(3):1–41, 2020.

Yuan Zhang, Jason Baldridge, and Luheng He. Paws: Paraphrase adversaries from word scrambling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 1298–1308, 2019.

Haiteng Zhao, Chang Ma, Xinshuai Dong, Anh Tuan Luu, Zhi-Hong Deng, and Hanwang Zhang. Certified robustness against natural language attacks by causal intervention. In *International Conference on Machine Learning*, pp. 26958–26970. PMLR, 2022.

Junbo Zhao, Yoon Kim, Kelly Zhang, Alexander Rush, and Yann LeCun. Adversarially regularized autoencoders. In *International conference on machine learning*, pp. 5902–5911. PMLR, 2018.

Zhengli Zhao, Dheeru Dua, and Sameer Singh. Generating natural adversarial examples. *arXiv preprint arXiv:1710.11342*, 2017.

Wanjun Zhong, Duyu Tang, Zenan Xu, Ruize Wang, Nan Duan, Ming Zhou, Jiahai Wang, and Jian Yin. Neural deepfake detection with factual structure of text. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 2461–2470, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.193. URL https://aclanthology.org/2020.emnlp-main.193.

Xin Zhou, Ruotian Ma, Yicheng Zou, Xuanting Chen, Tao Gui, Qi Zhang, Xuan-Jing Huang, Rui Xie, and Wei Wu. Making parameter-efficient tuning more efficient: A unified framework for classification tasks. In *Proceedings of the 29th International Conference on Computational Linguistics*, pp. 7053–7064, 2022.

Chen Zhu, Yu Cheng, Zhe Gan, Siqi Sun, Tom Goldstein, and Jingjing Liu. Freelb: Enhanced adversarial training for natural language understanding. *arXiv preprint arXiv:1909.11764*, 2019.

A Appendix

TextFooler: Generated Examples

A random selection these attacks on round 1 of the ANLI dataset are included in Table 9. A random selection of these attacks on the first round 2 (Analogous to round 1 in ANLI) of the Hate Speech Detection dataset are included in Table 13. Trends for this attack type are explained in the table captions.

BertAttack: Generated Examples

A random selection these attacks on round 1 of the ANLI dataset are included in Table 7. A random selection of these attacks on the first round 2 (Analogous to round 1 in ANLI) of the Hate Speech Detection dataset are included in Table 14. Trends for this attack type are explained in the table captions.

CT-GAN: Generated Examples

A random selection these attacks on round 1 of the ANLI dataset are included in Table 8. Trends for this attack type are explained in the table captions.

DI method: Generated Examples

A random selection these attacks on round 1 of the ANLI dataset are included in Table 10. A random selection of these attacks on the first round 2 (Analogous to round 1 in ANLI) of the Hate Speech Detection dataset are included in Table 15. Trends for this attack type are explained in the table captions.

ICE method: Generated Examples

A random selection these attacks on round 1 of the ANLI dataset are included in Table 11. A random selection of these attacks on the first round 2 (Analogous to round 1 in ANLI) of the Hate Speech Detection dataset are included in Table 12. Trends for this attack type are explained in the table captions.

Implementation Details for DI Method.

In this section we give more specifics needed to replicate the DI method. The general outline of the method is provided in Section 4.

For the generator, we use the default settings for the base T5 model, with no changes to the hyper-parameters (Raffel et al., 2020b)).The base T5 model is then trained on ANLI R1 attacks, with the premise and the label as the input (delimited by a special token [NLI_LABEL]:). The label is denoted by one of the three letters: (e, n, c) to denote entailment, neutral, and contradiction. The output sequence that the T5 model is trained to decode is the attack hypothesis in the ANLI R1 data. We use the standard cross-entropy loss function as described in the base T5 model and use the train split of the ANLI R1 data. We choose the best checkpoint based on the token-level accuracy in generating the hypotheses in the validation set of the ANLI R1 dataset. We then produce an in-sample generated attack dataset by producing multiple hypotheses (up to 100) on the train split of ANLI R1. Thus, we ensure that there is no leakage of attacks that were not already previously seen by the base classifier.

The cross-entropy loss of the generated text probabilities $\widehat{p_{\mathbf{a}_0}}(x_i, y)$ and the target text x_o, assuming a maximum sequence length K, and vocabulary $\mathbf{V}$ is given by:

$$\mathcal{L}_{ce}(x, y) = -\sum_{k=1}^{K} \sum_{v \in \mathbf{V}} \mathbb{1}_{x_{o,k}=v} \log(\widehat{p_{\mathbf{a}_0}}(x_i, y)_{k,v}) \tag{2}$$

For the classifier, we use the BERT-Large pre-trained model, and first fine-tune it on the MNLI+SNLI data mixture as per the ANLI benchmark. We then fine-tune that classifier further on the ANLI R1 train

Original ANLI R1 Premise	Original ANLI R1 Hypothesis ($\mathbf{a}_0$)	BertAttack(R1): Generated Hypothesis ($\mathbf{a_g}$)	Original Label	Label Still Correct?	Base ($\widehat{y}_{\theta_0}$)	Base + R1 ($\widehat{y}_{\theta_1}$)	Base + R1 + R2 ($\widehat{y}_{\theta_2}$)
The Other One is the third solo album by former Fleetwood Mac guitarist Bob Welch. The track "Future Games" was first released on the Fleetwood Mac album of the same name in 1971. Members of Welch's backing band also make songwriting contributions here though the majority of tracks are Welch's own	Members of Welch's backing band include Maynard James Keenan who assisted Bob Welch with almost every track on The Other One."	Member of Welch's backing band encompass Menard James Keenan who helped Boba Welsh with about everything trails on The Else One	e	n	n/a	n/a	n/a
The Walloon Legion (French: "Légion Wallonie") was a collaborationist volunteer unit recruited from Belgium's French-speaking population in Wallonia and Brussels during the German occupation of World War II. The Walloon Legion served in the Wehrmacht, later in the Waffen-SS, on the Eastern Front on both front line and reserve duties.	The Walloon Legion was partly recruited from Brussels.	The Walloon Legion was not a volunteers units.	e	n	n/a	n/a	n/a
Irma Pezzia Haubold (November 20, 1908 – April 4, 1996) was an American artistic gymnast. She competed at the 1936 Summer Olympics and placed fifth with the team. She was married to a fellow Olympic gymnast Frank Haubold. They were the first married couple of compete in the same Olympics.	Irma competed in the 1940 Summer Olympics.	Irma participated in the 1940 Hsia Olympics.	c	y	e	c	c
Mia Foni is the debut album of Greek American singer Annet Artani. It features 19 tracks in both Greek and English, including "Why Angels Cry", the song that Annet performed in the Eurovision Song Contest 2006 in Athens representing Cyprus. The album was released in both Greece and Cyprus where it entered the top 10.	Annet Artani's album have 15 tracks.	Annet Artani's albums have 15 trajectories.,	c	y	e	e	c
Ernest Thompson Willows (1886–1926) was a pioneer Welsh aviator and airship builder. He became the first person in the United Kingdom to hold a pilots certificate for an airship when the Royal Aero Club awarded him "Airship Pilots Certificate No. 1	Ernest Thompson Willows was a pilot.	Ernesto Thomson Willows was a experimenter	n	y	c	c	n
James Hugh Sinclair (16 October 1876 – 23 February 1913) was a South African cricketer who played in 25 Tests from 1896 to 1911. He scored South Africa's first three Test centuries and was the first person from any country to score a century and take five wickets in an innings in the same Test. He is one of the fastest-scoring Test batsmen of all.	They refered to him as a fast scorer.	They refered to him as a rapids scorer.,	e	y	n	e	e

Table 7: Selected examples from BertAttack(R1) attacks. Like the TextFooler attacks in Table 9 this attack makes small modifications to the hypothesis, explaining the high level of distributional similarity to ANLI rounds 1 and 2. Unlike TextFooler, however, BertAttack makes more sophisticated contextual phrase substitutions. For example in row 2, "partly recruited from Brussels" is replaced with "not a volunteers units."

data split. We choose the checkpoint based on the best validation set accuracy of ANLI R1 dataset. This fine-tuned model forms the basis of our comparison, and our goal is to improve adversarial robustness beyond what this model can achieve.

In the DI method, we then fine-tune the Base + R1 model on the generated examples. For all fine-tuning processes, we used an adamw optimizer in Jax + Tensorflow. The following learning rate schedule was used: There are first 3,681 warm-up steps at the initial learning rate of 3.0e-05. Then for the next 36,813 steps the learning rate decays at a linear rate (Though we only fine-tune for 40k steps total). The checkpoint that performs best on the validation split is selected for the next stage.

Original ANLI R1 Premise	Related Original ANLI R1 Hypotheses ($\mathbf{a}_0$)	CT-GAN(R1): Generated Hypothesis ($\mathbf{a_g}$)	Original Label	Label Still Correct?	Base ($\widehat{y}_{\theta_0}$)	Base + R1 ($\widehat{y}_{\theta_1}$)	Base + R1 + R2 ($\widehat{y}_{\theta_2}$)
The Parma trolleybus system (Italian: "Rete filoviaria di Parma") forms part of the public transport network of the city and "comune" of Parma, in the region of Emilia-Romagna, northern Italy. In operation since 1953, the system presently comprises four urban routes.	"Since 1953 the Parma Troleybus system has been comprised of four urban routes."	Parma has been in operation for the entire region of Emilia-Romagna, northern Italy.	e	y	e	e	e
Spoons was a comedy sketch show first broadcast on the United Kingdom's Channel 4 from 30 September 2005. In the United States, "Spoons" is broadcast on BBC America. The relationship themed show combined recent trends in sketch shows—dark content, strong language, and recurring catchphrases.	Spoons the comedy sketch show first appeared on the United Kingdom's Channel 4 a year after 2005.	Spoons was a comedy sketch show first broadcast on the United Kingdom's Channel 4 from 30 September 2005	c	n	n/a	n/a	n/a
The 1992 Boise State Broncos football team represented Boise State University in the 1992 NCAA Division I-AA football season. The Broncos competed in the Big Sky Conference and played their home games at Bronco Stadium in Boise, Idaho. Led by sixth-year head coach Skip Hall, Boise State finished the season 5–6 overall and 3–4 in conference.	["After 1991 Boise State Broncos competed in the Big Sky Conference.", "Boise State won four games at their home stadium in 1993.]"	The 1992 Boise State Broncos football team competed in the Big Sky Conference and played their home games at Bronco Stadium.	e	y	e	e	e
Other ranks (or ORs) in the Royal Danish Army, Royal Danish Navy and Royal Danish Air Force is a joined term for military personnel that are not officers of various forces, by the NATO system of ranks and insignia. These personnel are NCOs and privates, seamen or aircraftsmen	There are no NCOs, non-commissioned officers, in the Danish Military services.	These personnel are NCOs and privates, seamen or aircraftsmen.	c	n	n/a	n/a	n/a
"Drunk Girls" is a song by American rock band LCD Soundsystem. It was released as the second single from their third studio album, "This Is Happening" (2010), on May 3, 2010. Band frontman James Murphy has described the song as "dumb" but added "I like dumb, short stuff." The 7" single features a cover of the song by San Francisco psychedelic rock band Wooden Shjips.	James Murphy has said "I like dumb, short stuff".	"Band frontman James Murphy has said "I like dumb, short stuff".	e	y	e	e	e
Viru is a 5.0% ABV pilsner-style beer brewed in Estonia. It is brewed in the country's second largest city, Tartu, by the A. Le Coq brewery. The brand is owned by Baltic Beer Company Ltd (formerly Brand Independence Ltd), based in London, UK, and is brewed under licence in Estonia. A. Le Coq is the second largest brewery in Estonia, with a market share of 36.8% in 200	Viru is a 5.0% ABV pilsner-style beer brewed in Estonia. It is brewed in the country's second largest city, Tartu, by the A. Le Coq brewery. Lately there have been talks of setting up another brewery in the same region	Viru is brewed in the country's second largest city, Tartu, by the A. Le Coq brewery	n	n	n/a	n/a	n/a

Table 8: Selected examples from CT-GAN(R1) attacks. These attacks most often are direct sub-phrases of the premise. While they generate the entailment class examples correctly, almost no new information is ever added. Meanwhile, the tendency to pick up sub-phrases from the premise means the other two classes, neutral and contradiction, are largely generated incorrectly (See row 2 and 4). Further, since these incorrect example are often a verbatim copy of the premise, we may be teaching the model to not mark entailment for content directly in the premise).

Implementation Details for ICE Method.

In this section we give more specifics needed to replicate the ICE method. The general outline of the method is provided in Section 4.

Original ANLI R1 Premise	Original ANLI R1 Hypothesis ($\mathbf{a}_0$)	TextFooler(R1): Generated Hypothesis ($\mathbf{a_g}$)	Original Label	Label Still Correct?	Base ($\widehat{y}_{\theta_0}$)	Base + R1 ($\widehat{y}_{\theta_1}$)	Base + R1 + R2 ($\widehat{y}_{\theta_2}$)
"Aama" (Literally: Mother) is the first Nepali movie produced in Nepal, starring Shiva Shankar Manandhar and Bhuvan Chand (Thapa) as the leading actors. The movie was produced by the Information Department of the Nepalese Government and released on October 7, 1964. Bollywood film maker Hira Singh Khatri was invited by the late King Mahendra to direct the first Nepali movie.	Hira Singh Khatri was the director for the movie "Aama".	Hira Seng Khatri was the headmaster for the film "Aama".	"e"	False	n/a	n/a	n/a
"Dangerous" is a song by American electronic music project Big Data, from their debut EP "1.0" (2013) and their debut studio album "2.0" (2015). It features American indie rock band Joywave, with vocals being performed by the band's lead singer Daniel Armbruster.	Big Data did not produce any albums in 2007.	Massive Data did not engender any albums in 2007.	"e"	False	n/a	n/a	n/a
Camp Al-Saqr, referred to by some media sources as Camp Falcon, Forward Operating Base Falcon, Joint Service Station (JSS) Falcon, or Combat Outpost Falcon, was a United States military forward operating base in Iraq a short distance outside Baghdad, some 13 km south of the Green Zone. In OIF 2004; it was designated as "Camp Ferrin-Huggins". s of 2009 , the base housed up to 5,000 troops.	None of the soldiers agreed with the names given to the camp.	None of the soldiers countersigned with the names gave to the campground.	"n"	True	"e"	"e"	"e"
Wolf hunting with dogs is a method of wolf hunting which relies on the use of hunting dogs. While any dog, especially a hound used for hunting wolves may be loosely termed a "wolfhound", several dog breeds have been specifically bred for the purpose, some of which, such as the Irish Wolfhound, have the word in their breed name.	wolves are hunting dogs	wolfe are hunting dogs	"c"	True	"e"	"n"	"e"
The discography of American metalcore band As I Lay Dying consists of 6 studio albums, 2 compilation albums, 1 video album, 11 singles and 15 corresponding music videos as well as 1 split album with fellow metalcore band American Tragedy called "As I Lay Dying/American Tragedy".	The bands As I Lay Dying has filmed over 5 dozen music videos.	The strips As I Lay Death has videotaped over 5 dozen music videos.	"c"	False	n/a	n/a	n/a
Sir John Peebles Arbuthnott, PPRSE, FRCPSG, FMedSci, FRCPath (born 8 April 1939) is a Scottish microbiologist, and was Principal of the University of Strathclyde. He succeeded Lord Wilson of Tillyorn as President of The Royal Society of Edinburgh in October 2011 and was succeeded by Dame Jocelyn Bell Burnell in October 2014.	Sir John Peebles started out as a teacher.	Monsieur John Peebles commenced out as a teacher.	"n"	True	"e"	"e"	"e"

Table 9: Selected examples from TextFooler(R1) attacks. These attacks make small random modifications to the hypothesis, explaining the high level of distributional similarity to ANLI rounds R1 and R2. More examples are included in Supplemental Materials.

In addition to the DI method setup, we train a linear classifier using the T5 generator. The linear classifier takes in each hidden layer at the end of each transformer unit. We flatten and concatenate the hidden layers before feeding them into a single dense layer + layer norm with 3 as the output dimension size (one for each class in NLI). This is a departure from the PPLM implementation which sums the hidden layers (Dathathri et al., 2020). Adding another layer, even a small 3 by 3 dense layer, does help improve this linear classifier but can slow down the attack generation process. So, we did not use any additional layers. The linear

classifier still does not perform as well as T5 or BERT-Large on the NLI task (71% accuracy on MNLI dev as compared to 91% for T5 and 89% for BERT-Large).

Assuming that the set of labels $\mathbf{m}$: {entailment, neutral, contradiction}, the classifier optimizes the following cross-entropy classification loss:

$$\mathcal{L}_c(x, y) = -\sum_{m \in \mathbf{m}} y \cdot \log(\texttt{LC}(H(x))_m) \tag{3}$$

Further, the reconstruction loss used to warm-start the generator is given by:

$$\mathcal{L}_r(x, y) = -\sum_{k=1}^{K} \sum_{v \in \mathbf{V}} \mathbb{1}_{x_{o,k}=v} \log(\widehat{p_{\mathbf{a}_0}}(x, y)_{k,v}) \tag{4}$$

During attack generation, we use a high α beam parameter, 0.7 for ANLI, and 0.8 for the Toxicity dataset. For additional example diversity, the weight on the reconstruction loss can be made negative: results here are presented with a reconstruction loss weight of 0 for ANLI, and -1.5 for the Toxicity dataset. On each step that we update the T5 parameters, we smooth the update (final parameter = original parameter * 0.75 + new parameter * 0.25). We did not substantially explore these hyper-parameters, as each example generation run takes $\sim$ 100 GPU hours from warm-up steps to example generation.

We only include generated attacks if the linear classifier labels the generated example belongs to the original class. For example, we filter attacks that contain words in the set { "e", "n", "c" }, or more than two of each of those class labels (filtering "eee", or "cccc"), as these are NLI labels and an artifact of the NLI pretraining. We also filter attacks that exactly match a hypothesis in the observed set of examples, attacks that are less than two words long, or attacks that repeat a string of four characters more than four times (to avoid artifacts like: "cococococococococococococo" that may appear).

If the ICE method is used to generate data from a new round, (ex: R1 for ANLI), the base model is trained on the new round and synthetic data all in the same fine-tuning period (randomly shuffling the data). The original data is repeated until it is equal in number to the synthetic data. For example, if there are $1k$ original examples, and $5k$ generated data, the original set of $1k$ will be repeated five times, and the final shuffled dataset will have $10k$ examples. This ratio and shuffling method could be tuned further, but we did not attempt this.

Implementation Details for CT-GAN.

To implement the CT Generative Adversarial Network, adapted to the ANLI task, we use the TF-GAN library [1] with a generator and discriminator initialized with the T5 and the BERT-Large model fine-tune on MNLI, SNLI, ANLI R1 data in a sequential manner. These initializations are inline with what we use in our DI and ICE methods. Then we train the generator and discriminator in an iterative manner. In the first step, we fine-tune the generator to generate examples similar to ANLI R1 and produce a misclassification error from the classifier using a modified min-max objective. In the next step, we further fine-tune the classifier on those generated examples, with the original label as that of the corresponding ANLI R1 examples. We then repeat these two steps 10 times. In each step, we train the generator for 20,000 steps, and the discriminator for 10,000 steps.

Implementation Details for TextFooler.

To generate text attacks based on word-substitution, we use the TextFooler recipe outlined in (Jin et al., 2020). We use the implementation in the TextAttack library [2] and produce augmentation examples by perturbs the ANLI R1 hypothesis through word-embedding swap. We generated 50 transformations per example, inline with the amount of data generated using our imitation and exploration based methods.

[1] https://blog.tensorflow.org/2019/08/introducing-tf-gan-lightweight-gan.html
[2] https://github.com/QData/TextAttack

Implementation Details for fine-tuning BERT Large.

We used the default parameters for the uncased BERT Large model detailed in the corresponding paper, (Devlin et al., 2019). The checkpoint that performed best on the ANLI dev set was used.

Human Rating guidelines.

The paper authors are the human raters. These are the guidelines we follow for edge cases.

- The current year is assumed to be 2019, since we are working with the base T5 model trained in that year, and our dataset, ANLI is also generated in that year.
- Examples with incorrect grammar are not penalized if the matter asserted fit the correct label. (ex: "They is a very popular band")
- Mild spelling mistakes, where the meaning is absolutely clear after reading the premise are not penalized: (example: FeliCa is a intelligent card system developed for utilizing in *electron* cash cards but has not yet gain endorsement to be utilizing in the United States).
- Spelling mistakes in named entities create a new entity (ex: "Kënga" is not the same as "K??nga" and "Mars" is not the same as "Mar").
- Word substitutions from other languages are okay if they are commonplace English knowledge (ex: "sir" –> "monsieur"), but not okay if they change the meaning of the sentence (ex: "life" –> "vie", which has an alternative English meaning).
- Repeating phrases is okay. (ex: Mark Donovan best known for his role in productions such as "Shaun of the Dead", "Shaun of the Dead", "Black Books", "In Bruges", and "Murder Investigation Team.") since they don't change the truth of the matter asserted.
- Tautologies are still considered "entailment" if they are true given the premise. (Ex: "The movie, The Golden Compass, is a movie.")

Additional ANLI Result: Holding out R3 as a future round and training on R1 and R2 yields similar conclusions

Table 17 extends the setting from Table 2 where R1 and R2 are past observed attacks available for training, and R3 is the held out set of future human attacks. All conclusions are the same as the ones drawn from Table 2.

Additonal Result: Size of $\mathbf{a_g}$ used to train $\widehat{y}_{\theta_1}$: More examples are better up until a point.

Here, we seek to understand the role of the number of generative adversarial examples used for fine-tuning the base classifier. We use all available ANLI R1 real adversarial examples for fine-tuning the adversarial generator, and decode the specified number of examples from the generator for fine-tuning the base classifier. We demonstrate that as the number of generated examples increase, adversarial robustness increases up to a point (up to 50x: 50 generated examples for every real adversarial example). After that point, we observe more generated examples impacts adversarial robustness negatively as shown in Table 16.

For the 50x experiment, however, we observe the best checkpoint while training BERT-Large is usually reached when the model has trained on only $\sim 30\%$ of the generated examples provided. Therefore these findings are inconclusive. This behavior could be due to the interaction between (1) increased label noise/diversity of examples generated as we increase the number of generated examples due to a quirk of the decoding process, and/or (2) the improvement in adversarial robustness due to more examples incorporated into training until overfitting to the generated distribution.

Sanity Check: More fine-tuning on real adversaries does not help.

We study whether the gains obtained by the generated examples could have been obtained otherwise merely by training longer on the ANLI R1 data itself. In Table 18, we see that if we had continued for more steps training on R1 data, we do not observe higher validation set accuracy as compared to training on training on DI(R1) generated attacks. (These figures are not directly comparable to 2 in the Results section of the main paper, since here we present validation set accuracy, and the Results section contains test set accuracy).

Original ANLI R1 Premise	Original ANLI R1 Hypothesis ($\mathbf{a}_0$)	DI(R1): Generated Hypothesis ($\mathbf{a_g}$)	Original Label	Label Still Correct?	Base ($\widehat{y}_{\theta_0}$)	Base + R1 ($\widehat{y}_{\theta_1}$)	Base + R1 + R2 ($\widehat{y}_{\theta_2}$)
The Coward is a 1915 American silent historical war drama film directed by Reginald Barker and produced by Thomas H. Ince. Ince also wrote the film's story and scenario with C. Gardner Sullivan. The film stars Frank Keenan and Charles Ray. John Gilbert also appears in an uncredited bit part. A copy of "The Coward" is preserved at the Museum of Modern Art.	The Coward is a part of American history and located in a museum for preservation.	The Coward is a 1915 American silent historical war drama film The Coward is a 1915 American silent historical war drama film	"e"	True	"c"	"c"	"c"
"Dangerous" is a song by American electronic music project Big Data, from their debut EP "1.0" (2013) and their debut studio album "2.0" (2015). It features American indie rock band Joywave, with vocals being performed by the band's lead singer Daniel Armbruster.	Big Data did not produce any albums in 2007.	It is unlike any other Electronic Music release, but the Joywave may have been proud of us with our music.	"e"	False	n/a	n/a	n/a
The Joint Aviation Authorities (JAA) was an associated body of the ECAC representing the civil aviation regulatory authorities of a number of European States who had agreed to co-operate in developing and implementing common safety regulatory standards and procedures. It was not a regulatory body, regulation being achieved through the member authorities.	The JAA is not a regulatory body.	authorities of a number of European states are from civil aviation authorities and represent the civil aviation authorities of a number of European States who had agreed to regulate these standards and safety regulations.	"e"	False	n/a	n/a	n/a
11 Flowers () is a 2011 film directed by Wang Xiaoshuai centered on a boy named Wang Han who loses his white shirt which shows that he is his school's best gymnast. The loss of the shirt is to take on greater metaphorical meaning as the film progresses which is inspired by the director's own experience as a youth during the cultural revolution and the more general confusion of childhood.	Wang Xiaoshuai directed 11 films in 2011.	The loss of the shirt is to take on great film productions as the film progresses when it progresses to the end of the film.	"n"	False	n/a	n/a	n/a
Tiffanie DeBartolo (born November 27, 1970) is an American novelist, filmmaker, and co-founder of independent record label Bright Antenna. She has written two novels, "God-Shaped Hole" and "How To Kill a Rock Star". She wrote and directed the film "Dream for an Insomniac", featuring Jennifer Aniston, but which had a very limited release in 1996.	Jennifer Aniston was not an actress in any movie directed by Tiffanie DeBartolo.	The film "Dream for an Insomniac" written and directed by Tiffanie DeBartolo was widely released in 1994.	"c"	True	"n"	"n"	"n"
Viru is a 5.0% ABV pilsner-style beer brewed in Estonia. It is brewed in the country's second largest city, Tartu, by the A. Le Coq brewery. The brand is owned by Baltic Beer Company Ltd (formerly Brand Independence Ltd), based in London, UK, and is brewed under licence in Estonia. A. Le Coq is the second largest brewery in Estonia, with a market share of 36.8	A. Le Coq is the second largest brewery in Estonia, with a market share of 36.8% in 2005. It had an increased market share in 2015.	A. Le Coq is the second largest brewery in Estonia, with a market share of 36.8% in 2005. The A. Le Coq brewery makes a lot of beer in Latvia, which had a market share of 36.8% in 2005.	"n"	True	"e"	"e"	"e"

Table 10: Selected examples from DI(R1) attacks. These generated hypotheses deviate pretty significantly from the original hypothesis. They also incorporate knowledge from the premise (ex: "second largest" phrase is common to the premise and generated hypothesis in the Le Coq example on the bottom row). More examples are included in Supplemental Materials. Unlike the other methods, even when the label is incorrect, the generated hypothesis is usually still a pretty cogent sentence.

Original ANLI R1 Premise	Selected Original ANLI R1 Hypotheses for this premise ($\mathbf{a}_0$)	ICE(R1): Generated Hypothesis ($\mathbf{a_g}$)	Desired Label	Label Correct?	Base ($\widehat{y}_{\theta_0}$)	Base + R1 ($\widehat{y}_{\theta_1}$)	Base + R1 + R2 ($\widehat{y}_{\theta_2}$)
The Coward is a 1915 American silent historical war drama film directed by Reginald Barker and produced by Thomas H. Ince. Ince also wrote the film's story and scenario with C. Gardner Sullivan. The film stars Frank Keenan and Charles Ray. John Gilbert also appears in an uncredited bit part. A copy of "The Coward" is preserved at the Museum of Modern Art.	["Reginald Barker directed a movie in 1915.", "A 1915 American silent historical war drama film is preserved at the Museum of Modern Art.", $\cdots$]	Reginald Barker movie was a silent drama.	"e"	True	"c"	"c"	"c"
11 Flowers () is a 2011 film directed by Wang Xiaoshuai centered on a boy named Wang Han who loses his white shirt which shows that he is his school's best gymnast. The loss of the shirt is to take on greater metaphorical meaning as the film progresses which is inspired by the director's own experience as a youth during the cultural revolution and the more general confusion of childhood.	["The movie was inspired by the director's experience as a youth during the american revolution.", "11 Flowers () is a 2011 film directed by Wang Xiaoshuai about boy named Wang Han who loses his white shirt and is no longer a top gymnast. The shirt becomes a metaphor for childhood and growing up in China.", $\cdots$]	11 Flowers () is a 2011 film directed by Wang Xiaoshuai about boy named Wang Han who loses his white shirt and is top gymnast. The shirt becomes a metaphor for the director's experience of childhood. The shirt becomes a metaphor for the directors's experience of childhood.. The shirt becomes a metaphor for the film's story.	"e"	True	"c"	"c"	"c"
Domestically, he has played for Khan Research Laboratories, Lahore, Lahore Eagles, Lahore Lions, Lahore Qalandars, Pakistan A and Huntly (Scotland) during his career. He was the captain of Lahore Qalandars in the first edition of the Pakistan Super League.	["There are multiple teams based in Lahore, Pakistan.", "There has been at least one domestic sports league in Pakistan.", $\cdots$]	There are at least several teams in life	"n"	True	"n"	"n"	"n"
Benjamin A. Muncil (28 Aug 1867 – 16 Dec 1930) was an American master builder in the Adirondacks early in the 20th century. He was a major figure in the architectural development of the Adirondack Great Camps; among his many projects was Marjorie Merriweather Post's Camp Topridge, Northbrook Lodge, and White Pine Camp, a summer White House of US President Calvin Coolidge.	["Benjamin Muncil has built many things", "One of Muncil's projects was created for a former US President. ", $\cdots$]	One of the best American master builder in the Adirondacks projects built many things created for a former US President.	"e"	True	"e"	"e"	"e"
Piton is a Pilsner beer brand from the island of Saint Lucia, brewed by Windward & Leeward Brewing Limited, which is owned by Heineken. The beer was named for the Gros Piton and Petit Piton mountains on the island. It was first brewed on October 7, 1992.	["Even though the beer was first brewed in public in October, many tasted it previously", "Even though the beer was brewed in October, it is not certain where that occurred ", $\cdots$]	Even though the beer was brewed in public in October, no one tasted it previously the first of taste it formerly	"n"	False	n/a	n/a	n/a
Benjamin A. Muncil (28 Aug 1867 – 16 Dec 1930) was an American master builder in the Adirondacks early in the 20th century. He was a major figure in the architectural development of the Adirondack Great Camps; among his many projects was Marjorie Merriweather Post's Camp Topridge, Northbrook Lodge, and White Pine Camp, a summer White House of US President Calvin Coolidge.	["Benjamin Muncil has built many things", "One of Muncil's projects was created for a former US President. ", $\cdots$]	One of the best American master builder in the Adirondacks projects built many things created for a former US President.	"e"	True	"e"	"e"	"e"
Maxim "Max" Bemis ( ) is the lead singer, primary composer and primary lyricist of the band Say Anything. He also sings alongside Chris Conley in the supergroup Two Tongues (which features band members from Say Anything and Saves the Day), plays alongside his wife Sherri Dupree-Bemis under the name Perma, and is a writer for Marvel Comics.	['Maxim Bemis is a composer, lyricist, writer and a singer.", $\cdots$]	Maxim Bemis is not a composer, writer and singer.	"e"	False	n/a	n/a	n/a

Table 11: Selected examples from ICE(R1) attacks. This method generates hypotheses that learn from past seen hypotheses and the premise. Common failure modes include generating nonsensical sentences or getting

Original Hate Speech R2a Example($\mathbf{a}_0$)	Matched Hate Speech R2b Example($\mathbf{a}_0$)	ICE(R2): Generated Example ($\mathbf{a_g}$)	Desired Label	Label Correct?	Base(R1) ($\widehat{y}_{\theta_0}$)	Base(R1) + R2 ($\widehat{y}_{\theta_1}$)
are black folks extinct? cause they should be (hate)	are black panthers extinct? (not hate)	man black folks always wear black.	"h"	False	"h"	"h"
lesbian r subhummmmmman (hate)	Why the fuck do you think lesbians are subhuman? (not hate)	lesban facias doesn't deserve rights.	"h"	True	"h"	"h"
Who would cut off their wiener for a hole...dont get it. it's like my brain just farts when I try to think of what must go through these people's heads to ever contemplate making that sort of crazy choice (hate)	Who would cut their wiener in several pieces for lunch..dont get it. it's like my brain just farts when I try to think of what must go through these people's heads to ever contemplate making that sort of crazy choice. (not hate)	Who would cut off their armies?	"n"	True	"h"	"n"
I can believe that she thinks arabs are the most kind ethnic group, like come on they are lovely (not hate)	I cant believe that she think arabs are the most kind ethnic group, like come on more like least (hate)	how can he believe he cannot teach anyone so much English	"n"	True	"n"	"n"
black people are super clean (not hate)	negroes are not clean (hate)	Black people are BLIGGNOW	"n"	True	"h"	"h"
My work doesn't cater for people with invisible disabilities. (not hate)	Glad my work doesn't cater for people with disabilities. (hate)	My work doesn't cater for pet dogss who're close to pet parents.	"n"	True	"h"	"h"
look dude, gyps are evil and basically cant be trusted. Especially with their cheating crystal balls, they freak me the fuck out. Grubby looking caravan lovers. (hate)	look dude, they are evil and basically cant be trusted. Especially with their cheating crystal balls, they freak me the fuck out. Grubby looking witches. (not hate)	look dude, chinese climates are evil and basically cant be trusted Especially with chinese chia	"h"	False	"h"	"h"
your personality is so ugly it could use makeup, typical black chick (hate)	your personality is so ugly it could use makeup (not hate)	Your black for black for a black person is too heavy for whites	"h"	True	"h"	"h"
all women are not to be trusted, they are snakes, vicious vipers trying to trick their men (hate)	all men are not to be trusted, they are snakes, vicious vipers trying to trick their women (not hate*)	all women are bitches everyone they have all fucking wifes	"h"	True	"h"	"h"
the bathroom is scummy we need some viakal and to get the cleaning sqaud in here. Maybe I'll do that tomorrow, ima go sort out somet stuff now (not hate)	the bathroom is so scummy it looks like an Arab house, honestly. We need some viakal and to get the cleaning sqaud in here. Maybe I'll do that tomorrow, ima go sort out somet stuff now (hate)	the bathroom is scccccccummy - we need some viakale and to think there is anything wrong with us just shut up and think about it, there should be black people living there	"h"	True	"h"	"h"
Islamic Relief should be banned (not hate*)	Islamic Relief should be banned, so should islamic people (hate)	Islamic Relating To The Devil *	"h"	True	"h"	"h"

Table 12: Selected examples from ICE(R2) attacks on the Hate Detection Task. This method generates text that learn from past seen text. Common failure modes include generating nonsensical sentences (see: row 5) or not being able to correctly label text (see: row 1). Some of the generated text are very random compared to the original examples. Often irrelevant text is added (see: row 7). Some of these lift phrases from other examples (ex: "deserve rights" in row 2 comes from another example). Sometimes the original label itself is incorrect (marked with *).

Original Hate Speech R2a Example($\mathbf{a}_0$)	Matched Hate Speech R2b Example($\mathbf{a}_0$)	TextFooler(R2): Generated Example ($\mathbf{a_g}$)	Desired Label	Label Correct?	Base(R1) ($\widehat{y}_{\theta_0}$)	Base(R1) + R2 ($\widehat{y}_{\theta_1}$)
are black folks extinct? cause they should be (hate)	are black panthers extinct? (not hate)	are black folks missing? cause they should be	"h"	True	"h"	"h"
lesbian r subhummmmmman (hate)	Why the fuck do you think lesbians are subhuman? (not hate)	lesbian rs subhummmmmman	"h"	True	"h"	"h"
Who would cut off their wiener for a hole...dont get it. it's like my brain just farts when I try to think of what must go through these people's heads to ever contemplate making that sort of crazy choice (hate)	Who would cut their wiener in several pieces for lunch..dont get it. it's like my brain just farts when I try to think of what must go through these people's heads to ever contemplate making that sort of crazy choice. (not hate)	Who would chop off their wiener for a hole...dont get it. it's like my brain just farts when I try to think of what must go through these people's heads to ever contemplate doing that sorted of psychopath choice	"h"	True	"n"	"h"
I can believe that she thinks arabs are the most kind ethnic group, like come on they are lovely (not hate)	I cant believe that she think arabs are the most kind ethnic group, like come on more like least (hate)	I can believe that she idea arabs are the most kind ethnic group, like come on they are lovely	"n"	True	"h"	"n"
black people are super clean (not hate)	negroes are not clean (hate)	negro people are super clean	"n"	True*	"h"	"n"
My work doesn't cater for people with invisible disabilities. (not hate)	Glad my work doesn't cater for people with disabilities. (hate)	My work doesn't cater for people with invisible disability.	"n"	True	"h"	"n"
look dude, they are evil and basically cant be trusted. Especially with their cheating crystal balls, they freak me the fuck out. Grubby looking witches. (not hate)	look dude, gyps are evil and basically cant be trusted. Especially with their cheating crystal balls, they freak me the fuck out. Grubby looking caravan lovers. (hate)	look dude, they are wicked and basically didnt be trusted. Especially with their cheating crystal balls, they freak me the fuck out. Grubby looking witches.	"n"	False	"n"	"n"
your personality is so ugly it could use makeup, typical black chick (hate)	your personality is so ugly it could use makeup (not hate)	your personality is so ugly it could use makeup, typical black dame	"h"	True	"n"	"h"
all women are not to be trusted, they are snakes, vicious vipers trying to trick their men (hate)	all men are not to be trusted, they are snakes, vicious vipers trying to trick their women (not hate*)	all men are not to be trusted, they are serpents, vicious vipers trying to trick their women,	"h"	False**	"h"	"n"
the bathroom is scummy we need some viakal and to get the cleaning sqaud in here. Maybe I'll do that tomorrow, ima go sort out somet stuff now (not hate)	the bathroom is so scummy it looks like an Arab house, honestly. We need some viakal and to get the cleaning sqaud in here. Maybe I'll do that tomorrow, ima go sort out somet stuff now (hate)	the toilettes is scummy we need some viakal and to get the cleaning sqaud in here. Possibly I'll do that tomorrow, ima go sort out somet stuff now	"n"	True	"n"	"n"
Islamic Relief should be banned, so should islamic people (hate)	Islamic Relief should be banned (not hate***)	Moslems Relief should be banned, so should islamic people	"h"	True	"h"	"h"

Table 13: Selected examples from TextFooler(R2) attacks on the Hate Detection Task. As in the ANLI task, it makes random word substitutions. In this task, it emulates the style of creating the matched example pretty well, which is perhaps why we see AUC improvement within R2 in this task with this dataset (*Whether "negro" is pejorative is contextual, and this rating decision could go the other way. **We don't agree with this label, but it is consistent with the label choices in the dataset. ***This label comes from the dataset and may be incorrect.)

Original Hate Speech R2a Example($\mathbf{a}_0$)	Matched Hate Speech R2b Example($\mathbf{a}_0$)	BertAttack(R2): Generated Example ($\mathbf{a_g}$)	Desired Label	Label Correct?	Base(R1) ($\widehat{y}_{\theta_0}$)	Base(R1) + R2 ($\widehat{y}_{\theta_1}$)
are black folks extinct? cause they should be (hate)	are black panthers extinct? (not hate)	are negro folks disappeared? reason they should be	"h"	True	"h"	"h"
lesbian r subhummmmmman (hate)	Why the fuck do you think lesbians are subhuman? (not hate)	lesbian rs subhummmmmman	"h"	True	"h"	"h"
Who would cut off their wiener for a hole...dont get it. it's like my brain just farts when I try to think of what must go through these people's heads to ever contemplate making that sort of crazy choice (hate)	Who would cut their wiener in several pieces for lunch..dont get it. it's like my brain just farts when I try to think of what must go through these people's heads to ever contemplate making that sort of crazy choice. (not hate)	Whom ought slice off their wiener for a hole...whats obtains it. it's likes my brain just farts when I endeavour to thinking of what owes going through these people's chiefs to ever envisage doing that sorts of maniac opted	"h"	True	"h"	"h"
I cant believe that she think arabs are the most kind ethnic group, like come on more like least (hate)	I can believe that she thinks arabs are the most kind ethnic group, like come on they are lovely (not hate)	I dunno believe that she believing arabs are the most types racism grouping, like arrive on more like fewer	"h"	???	"h"	"h"
black people are super clean (not hate)	negroes are not clean (hate)	nigger people are fabulous clean	"n"	False	"n"	"h"
My work doesn't cater for people with invisible disabilities. (not hate)	Glad my work doesn't cater for people with disabilities. (hate)	My corporation doesn't cater for people with undetectable incapacity.	"n"	True	"h"	"n"
look dude, they are evil and basically cant be trusted. Especially with their cheating crystal balls, they freak me the fuck out. Grubby looking witches. (not hate)	look dude, gyps are evil and basically cant be trusted. Especially with their cheating crystal balls, they freak me the fuck out. Grubby looking caravan lovers. (hate)	glance guy, they are wicked and primarily becuase be trustworthy. Specially with their cheating crystal balls, they flipper me the fucking out. Grubby looking bruges.	"n"	???*	"n"	"n"
your personality is so ugly it could use makeup, typical black chick (hate)	your personality is so ugly it could use makeup (not hate)	your personality is so hideous it did utilise makeup, typical black dame	"h"	True	"n"	"h"
all women are not to be trusted, they are snakes, vicious vipers trying to trick their men (hate)	all men are not to be trusted, they are snakes, vicious vipers trying to trick their women (not hate*)	all men are not to be trustworthy, they are snakes, ferocious viper tempting to stratagem their mujer,	"h"	False**	"h"	"n"
the bathroom is scummy we need some viakal and to get the cleaning sqaud in here. Maybe I'll do that tomorrow, ima go sort out somet stuff now (not hate)	the bathroom is so scummy it looks like an Arab house, honestly. We need some viakal and to get the cleaning sqaud in here. Maybe I'll do that tomorrow, ima go sort out somet stuff now (hate)	the toilettes is scummy we needs some viakal and to obtain the cleansing sqaud in here. Likely I'll do that tomorrow, ima going sorted out somet thing now,	"n"	True	"n"	"n"
Islamic Relief should be banned, so should islamic people (hate)	Islamic Relief should be banned (not hate*)	Islamists Relief should be prohibiting, so should islamist people "h"	True	"h"	"h"	

Table 14: Selected examples from BertAttack(R2) attacks on the Hate Detection Task. As in the ANLI task, it makes contextual word substitutions. In this task, it emulates the style of creating the matched example pretty well, which is perhaps why we see AUC improvement within R2 in this task with this dataset. Since it seeks out word synonyms from the model it attacks, it may have learned that "black people" and the slur, "nigger," are used to refer to the same category of people; it makes that substitution incorrectly in row 5 (*This label comes from the dataset and may be incorrect. **We don't agree with this label, but it is consistent with the label choices in the dataset.)

Original Hate Speech R2a Example($\mathbf{a}_0$)	Matched Hate Speech R2b Example($\mathbf{a}_0$)	DI(R2): Generated Example ($\mathbf{a_g}$)	Desired Label	Label Correct?	Base(R1) ($\widehat{y}_{\theta_0}$)	Base(R1) + R2 ($\widehat{y}_{\theta_1}$)
are black folks extinct? cause they should be (hate)	are black panthers extinct? (not hate)	are black folks extinct? can anyone else now avoid living in our countryside with the extinction?	"h"	False	"h"	"h"
lesbian r subhummmmmman (hate)	Why the fuck do you think lesbians are subhuman? (not hate)	Why the fuck do you think lesbians are subhuman?	"h"	True	"n"	"n"
Who would cut off their wiener for a hole...dont get it. it's like my brain just farts when I try to think of what must go through these people's heads to ever contemplate making that sort of crazy choice (hate)	Who would cut their wiener in several pieces for lunch..dont get it. it's like my brain just farts when I try to think of what must go through these people's heads to ever contemplate making that sort of crazy choice. (not hate)	Who would cut their wiener in serious negroes for a hole...dont get it. it's like my brain just farts when I try to think of what must go through these people's heads to ever contemplate making that sort of crazy choice.	"h"	???*	"n"	"h"
I cant believe that she think arabs are the most kind ethnic group, like come on more like least (hate)	I can believe that she thinks arabs are the most kind ethnic group, like come on they are lovely (not hate)	I can certainly believe that she thinks arabs are the most kind ethnic group, like come on they are lovely	"h"	False	"h"	"n"
black people are super clean (not hate)	negroes are not clean (hate)	black people are unclean	"n"	False	"h"	"n"
My work doesn't cater for people with invisible disabilities. (not hate)	Glad my work doesn't cater for people with disabilities. (hate)	Glad my work doesn't cater for people with invisible disabilities, always exempt from the amount of disabilities I can told me about their activities to show off other things	"n"	False	"h"	"n"
look dude, they are evil and basically cant be trusted. Especially with their cheating crystal balls, they freak me the fuck out. Grubby looking witches. (not hate)	look dude, gyps are evil and basically cant be trusted. Especially with their cheating crystal balls, they freak me the fuck out. Grubby looking caravan lovers. (hate)	they are evil and basically cant be trusted. Especially with their cheating crystal balls, they freak me the fuck out. Grubby looking with their cheating crystal balls.	"n"	???*	"n"	"n"
your personality is so ugly it could use makeup, typical black chick (hate)	your personality is so ugly it could use makeup (not hate)	your personality is so ugly it could use makeup, typycal women do not deserve to be allowed to work with yellow neighbours	"h"	True	"h"	"h"
all women are not to be trusted, they are snakes, vicious vipers trying to trick their men (hate)	all men are not to be trusted, they are snakes, vicious vipers trying to trick their women (not hate**)	all black women are not to be trusted, they are serpents, vicious vipers trying to trick their women,	"h"	True	"h"	"h"
the bathroom is scummy we need some viakal and to get the cleaning sqaud in here. Maybe I'll do that tomorrow, ima go sort out somet stuff now (not hate)	the bathroom is so scummy it looks like an Arab house, honestly. We need some viakal and to get the cleaning sqaud in here. Maybe I'll do that tomorrow, ima go sort out somet stuff now (hate)	the bathroom is so scummy we need some viakal and to get the cleaning sqaud in here. Maybe I'll do that tomorrow, ima go sort out somet stuff now	"n"	True	"n"	"n"
Islamic Relief should be banned, so should islamic people (hate)	Islamic Relief should be banned (not hate**)	Islamic Relief is trash. We need Muslim like the retarded	"h"	True	"h"	"h"

Table 15: Selected examples from DI(R2) attacks on the Hate Detection Task. This method on this task is trained with pairs that have opposite labels (one pair hate, the other not hate). As a result, the model often generates text closer to the matched example than to the original (ex: rows 3 to 5) which results in very high label confusion (*Label is unclear because while a group is not explicitly referred to, "crystal ball" can be a dog whistle. **This label comes from the dataset and may be incorrect.)

No. of generated examples	ANLI R1	ANLI R2	ANLI R3
1k	36.5	31.6	36.2
10k	41.8	32.3	36.3
100k	42.9	36.5	37.0
200k	45.0	36.6	37.2
400k	45.1	36.9	37.4
850k	**48.2**	**39.1**	**40.1**
1.6M	43.9	37.9	38.7

Table 16: Impact of the number of generated adversarial examples from DI(R1) used for the fine-tuning the base classifier, on future rounds of attack in ANLI R2 and R3

Model	a_0: R1	a_1: R2	a_2: R3
$\widehat{y}_{\theta_2}$: Base + R1 + R2	52.7	41.6	38.4
$\hookrightarrow$ + DI (R1 + R2)	52.8	44.2	42.3
$\hookrightarrow$ + ICE(R1 + R2)	**55.2**	**48.1**	**42.8**

Table 17: Improvement on ANLI accuracy (%) when trained on attacks generated only from Rounds 1 and 2. The notation, DI(R1 + R2), refers to the method DI using R1 and R2 data to generate more examples.

step #	base+R1	DI(R1)
6,135	**40.8%**	-
12,270	37.5%	36.4%
18,405	33.3%	**42.2**%
24,540	37.0%	33.3%
30,675	38.2%	33.3%
36,810	38.5%	37.3%
42,945	-	41.5%

Table 18: Comparison of ANLI R1 validation set accuracy when training on additional steps on ANLI R1 data, as compared to using the best checkpoint from ANLI R1 data, and using that to further train on generated DI(R1) adversarial examples.

Model	ANLI R1	SNLI	MNLI(m)
Base	0.21	0.90	0.86
Base + R1	0.44	0.82	0.77
Base + R1 + TextFooler(R1)	0.24	0.13	0.07
Base + R1 + DI(R1)	0.48	0.74	0.69
Base + R1 + ICE(R1)	0.51	0.75	0.66

Table 19: Trade-off between adversarial and clean accuracy as measured on ANLI R1 and MNLI/SNLI test data

Dataset	Round	Train	Validation	Test
ANLI	1	16946	1000	1000
	2	45460	1000	1000
	3	100459	1200	1200
HateSpeech	1	8844	1091	1111
	2	7998	999	999
	3	7960	995	995
	4	8122	1015	1015

Table 20: Number of human adversarial examples in ANLI and Hate speech detection task, by split.