

Learning the Legibility of Visual Text Perturbations

Dev Seth¹ Rickard Stureborg¹ Danish Pruthi* Bhuwan Dhingra¹

¹ Duke University

{ds447, rs541, bd149}@duke.edu

Abstract

Many adversarial attacks in NLP perturb inputs to produce visually similar strings (‘ergo’ → ‘εrgo’) which are legible to humans but degrade model performance. Although preserving legibility is a necessary condition for text perturbation, little work has been done to systematically characterize it; instead, legibility is typically loosely enforced via intuitions around the nature and extent of perturbations. Particularly, it is unclear to what extent can inputs be perturbed while preserving legibility, or how to quantify the legibility of a perturbed string. In this work, we address this gap by learning models that predict the legibility of a perturbed string, and rank candidate perturbations based on their legibility. To do so, we collect and release **LEGIT**, a human-annotated dataset comprising the legibility of visually perturbed text. Using this dataset, we build both text- and vision-based models which achieve up to 0.91 F1 score in predicting whether an input is legible, and an accuracy of 0.86 in predicting which of two given perturbations is more legible. Additionally, we discover that legible perturbations from the **LEGIT** dataset are more effective at lowering the performance of NLP models than best-known attack strategies, suggesting that current models may be vulnerable to a broad range of perturbations beyond what is captured by existing visual attacks.¹

1 Introduction

To manage the increasing demand for content moderation—e.g., detecting spam or toxic/hateful content on online platforms—organizations have turned to machine learning solutions. In response, users often resort to manipulating text to evade detection, removal, or search. For instance, hateful comments often comprise of visually similar characters to avoid automatic filtering (Le et al., 2022).

* Work done while at Carnegie Mellon University.

¹Data, code, and models are available at <https://github.com/dvsth/learning-legibility-2023>.



Figure 1: Visual attacks in the wild. Examples of Twitter users manipulating their tweets to evade the platform’s ‘sensitive content’ detection algorithms.

Since people read text visually, the manipulated content can still be easily understood and harm its target audience. These attacks started with simple ASCII substitutions like he11o (colloquially referred to as “leetspeak”), but have evolved into complex manipulations utilizing characters from different Unicode scripts (Flamand, 2008; Raymond, 1996). Figure 1 shows two such examples.

Unlike computer vision where there is an established notion of what constitutes an imperceptible perturbation (typically defined via the ℓ_∞ distance), most perturbations in text are perceptible. However, as long as the perceptible manipulations remain *legible*, the message could have its intended effect. The legibility of a text is determined by whether or not a literate person can decipher the altered words. The degree to which a piece of text can be perturbed, while maintaining legibility, depends on a multitude of factors such as its context, similarity to the original content, the positions of the perturbations, the background knowledge of the reader, etc.

However, many adversarial attacks enforce legibility only loosely based on intuitions about the nature of the attacks, e.g., that changing 1-2 characters in a sentence does not impact its legibility (Belinkov and Bisk, 2018; Pruthi et al., 2019).

In this work, we instead propose to learn the legibility of visual perturbations, by developing text- and vision-based models trained on legibility annotations from human subjects. The current focus of research on adversarial attacks is to find minimal perturbations required to break NLP models, and several recent findings suggest that models remain brittle to such perturbations (Eger et al., 2019; Dionysiou and Athanasopoulos, 2021; Pruthi et al., 2019). In contrast, our work attempts to uncover the space of *all legible perturbations* that we need to defend against. Towards our goal of characterizing the limits of legibility of perturbed texts, we make the following contributions:

First, we crowdsource human judgments about the legibility of different perturbations: specifically, we show annotators two perturbed versions of the same word and ask them which one, if any, they find more legible. Our perturbation strategy considers substituting letters in the word with Unicode characters drawn from a large subset of the Basic Multilingual Plane covering over 100 scripts from around the world.² In total, we collect 30,320 annotations, one each for 14,643 and 3,332 instances in the training and validation sets, respectively, and three each for the 4,113 instances in the test set. Using these preferences, we define a *pairwise legibility ranking task* as well as a *binary legibility classification task*. While the former allows making inferences about which candidate perturbation is *most* legible, the latter allows filtering out illegible perturbations altogether. For each task, we identify a *hard* subset of the collected data, which includes fine-grained comparisons expected to be more challenging for annotators and models alike.

Second, we use the labeled data to train models which predict the degree of legibility of a perturbed text. Specifically, we fine-tune pretrained vision (TrOCR; Li et al., 2021) and text-based (ByT5; Xue et al., 2022) models on the ranking and classification tasks. We find that TrOCR trained in a multi-task setup on both tasks achieves the best performance with 0.91 F1 score on the classification task and 0.86 accuracy on the ranking task.

²We consider 12,287 Unicode characters from codepoints 0x0000 to 0x2fff.

Interestingly, we find that the purely text-based ByT5 also achieves competitive performance on the classification task with 0.89 F1, suggesting that its pretrained byte representations encode aspects of visual similarity between Unicode characters. Further, we find that models have high F1 scores on the subset of data with high inter-annotator agreement: TrOCR achieves a 0.96 F1 score on test cases where all three annotators agree. We also note that legibility is a complex phenomenon—it doesn’t correlate trivially with the distance of the perturbation from the original text or the number of letters substituted.³

Third, we consider a word-level *perturbation recovery* task, which involves inferring the original word from its perturbed version. We evaluate GPT-3 (Brown et al., 2020) on this task, comparing its performance on legible perturbations from our perturbation strategy versus those generated by VIPER, a *Visual PERTurber* method proposed by Eger et al. (2019). We find that GPT-3 has a lower accuracy in recovering perturbations from our perturbation strategy, despite VIPER providing no guarantees on legibility. Additionally, we apply our findings to the important task of toxicity classification. We perturb a subset of the dataset using our perturbation strategy and find that it degrades the SOTA Detoxify (Hanu and Unitary team, 2020) classifier more than existing VIPER attacks. These findings demonstrate that existing attacks do not comprehensively cover the space of legible perturbations that can degrade model performance.

2 Related Work

Adversarial Attacks for NLP. A challenge in defining adversarial examples for text lies in characterizing the space of *equivalent* inputs to a training or test example which preserves the target label. While early work focused on adding distracting text to fool question answering systems (Jia and Liang, 2017), recent work utilizes more general strategies applicable to many tasks (Li et al., 2019; Morris et al., 2020; Jin et al., 2020). Many of these can be categorized as *word-level* synonym substitutions (Alzantot et al., 2018; Garg and Ramakrishnan, 2020; Li et al., 2020), or *character-level* legibility-preserving substitutions (Ebrahimi et al., 2018; Pruthi et al., 2019). Most attacks in either category are perceptible in that readers of

³A logistic regression model using these as features only agrees 56.7% of the time with authors’ legibility assessment.

the text can identify that it has been transformed, except for one notable exception where invisible characters and near-identical characters are used to render strings indistinguishable from the original (Boucher et al., 2022). Attacks based on visual similarity of characters have also been previously considered by Eger et al. (2019) who propose three attack strategies: ICES (based on rendered glyph similarity), DCES (based on bag-of-words textual similarity of Unicode codepoint descriptions), and ECES (based on adding diacritics to base characters). For ICES, they compute similarity by comparing raw pixel values of the renderings, which we improve upon here by utilizing a pretrained Optical Character Recognition (OCR) model. This produces a ‘smarter’ set of visual neighbors: e.g., mirror images of letters, scaled versions of letters (like O vs °) etc., which go beyond simple accents or modifiers. We also report in-depth comparisons between our perturbation strategy and the ECES and DCES approaches in section 5.

Legibility of Perturbed Inputs. Among character-level perturbation attacks, legibility has only been loosely enforced based on intuitions about the nature and the degree of manipulations. This often results in conservative substitutions which only represent a lower bound on the space of all legible perturbations. For instance, Pruthi et al., 2019 limit the attack to only 1-2 character changes (e.g., substitutions, deletions or additions) per input example; similarly, Ebrahimi et al., 2018 propose an attack strategy which specifically minimizes the number of character manipulations required in order to render the output legible. Attacks based on visual similarity usually constrain their attack surface to inputs which are above a threshold similarity (in pixel or embedding space) to the original input (Eger et al., 2019; Eger and Benz, 2020; Dionysiou and Athanasopoulos, 2021). In this work, by contrast, we directly address the question of what constitutes legible perturbations, with the aim of learning a grounded definition of legibility rather than assuming one *a priori*.

3 Legibility Tests

We adopt a supervised learning approach for determining the legibility of perturbed texts. In this section, we describe the process used for collecting the LEGIT dataset (which stands for LEGIBILITY Tests) and in the next section we describe the modeling techniques used for predicting the legibility

score and ranking different candidate perturbations.

Our setting involves one-to-one character substitutions at the word level, i.e., given a word (and no other context), we consider perturbations where each letter in the word may be replaced by a Unicode codepoint in 0x0000-0x2fff. Moreover, the substitutions are mutually independent and do not depend on the context of the other letters.

3.1 Perturbation Process

To generate perturbations for the data labeling task, we replace a subset of characters in a word with visually similar counterparts. Specifically, given a word w , we first randomly select a fraction $n \in [0, 1]$ of characters in that word to corrupt. Then, each of the chosen characters is replaced by its nearest neighbor at rank k in the embedding space generated by a model \mathcal{M} which encodes characters into visual features. Hence, there are three parameters involved in the perturbation process $\phi = \{n, k, \mathcal{M}\}$.

We experiment with several models to encode characters into visual features, all based on renderings of the Unicode codepoints into images. To keep visual representations consistent across models, we use GNU Unifont, rendering each glyph separately in 144px font size with black color, on a 224×224 px white background.⁴ Given the rendering, we compare 5 models to encode the features. Three are transformer-based: TROCR (‘base’) (Li et al., 2021), CLIP (‘vit-base-patch32’) (Radford et al., 2021), and BEiT (‘base-patch16-224-pt22k-ft22k’). One employs convolutional as well as transformer networks: DETR (Carion et al., 2020). The fifth model is a simple baseline: IMGDOT, which uses the (flattened) bitmap of a rendered character as its embedding vector. In preliminary experiments, 400 perturbed pairs were generated, with each pair using the same settings for k, n but using different models. The authors then independently ranked perturbations each pair based on their legibility. DETR- and BEiT-generated perturbations were ranked above other models’ perturbations 23% and 41% of the time, respectively, whereas CLIP and IMGDOT perturbations were preferred over others in 66% and 73% of cases. Hence, DETR and BEiT were excluded from further experiments. TROCR was included later, after verifying that it was preferred $\approx 50\%$ of the time against both CLIP and IMGDOT.

⁴Glyphs were rendered by the Pillow library (Clark, 2015).

For each of the chosen models, we compute the pairwise cosine distances between the model’s embedding vectors for all Unicode codepoints in the range 0x0000–0x2fff (excluding invalid or empty codepoints), and use these distances to find the nearest-neighbors for each character. Then, to perturb a given word w using the parameters $\phi = \{k, n, \mathcal{M}\}$, we first pick $\lfloor n|w| \rfloor$ characters uniformly at random to replace. For each character, we fetch its k -th nearest neighbor from the model \mathcal{M} . Finally, we apply these substitutions to the target word to obtain the perturbed word.

3.2 Pairwise Comparisons

We crowdsource legibility annotations for the perturbed words using Amazon’s Mechanical Turk. We collect annotations on both absolute legibility as well as relative preference between two differently perturbed inputs. Since annotators tend to produce higher quality annotations when comparing items rather than assigning absolute values (Callison-Burch et al., 2007; Liang et al., 2020), we design an annotation interface based on *pairwise comparisons* of two perturbed versions of the same word (Appendix A). Specifically, annotators see perturbations w_1, w_2 side-by-side, with the original word w hidden. They are asked to indicate which perturbation they find more legible by selecting exactly one of these four labels:

- L1:** w_1 is preferred
- L2:** w_2 is preferred
- BL:** both w_1, w_2 are equally legible
- NL:** neither w_1 nor w_2 is legible

L_1 and L_2 capture not only relative preferences between the two perturbations (used for the ranking task), but also indicate that the preferred perturbation is legible. However, these labels do not give us any information about the non-preferred perturbation. On the other hand, the BL (Both Legible) and NL (Neither Legible) options do not give us a ranking between the two words, but inform us about the legibility (or illegibility) of both words. In the next section, we use these labels to derive datasets for both a pairwise ranking task and a binary classification task.

We generate the data for annotation from English words consisting of the top 10,000 frequent words (as per Kaufman (2012)) in the Trillion

Word Corpus (Brants and Franz, 2006). We filter this vocabulary to remove words with lengths less than 4 or greater than 14, ending up with 7600 words. These words are randomly split into the train (65%), validation (15%), and test (20%) sets; all future perturbation pairs (w_1, w_2) generated for word w are added to the corresponding set, and the same sets are used for all experiments. To perturb a word w into the pair w_1, w_2 a model \mathcal{M} is picked at random from {TROCR, CLIP, IMGDOT} (the three best models from our initial perturbation analysis). We sample $k \sim \mathcal{N}(\mu_k, \sigma_k^2)$ and similarly for n , applying the appropriate bounds to keep $k > 0$ and $n \in [0, 1]$. The initial values are $\mu_k = 25, \sigma_k^2 = 10, \mu_n = 0.5, \sigma_n^2 = 0.2$.

3.3 Adaptive Annotations

The space of all possible perturbations of a word is vast, and sampling the parameters ϕ based on the priors above is unlikely to yield difficult perturbations which lie at the boundary of legibility. In order to identify such perturbations, we collect data over multiple rounds using an *adaptive* process for generating the pairs. In the first round, the pairs are generated as described above and annotated by the crowd-workers. In the following rounds, pairs are generated taking into account the last round of annotations. Specifically, the ϕ_1, ϕ_2 for each successive round are chosen to make the next round of labeling harder for annotators. This is accomplished by manipulating the Gaussian used to generate k, n , i.e. by shifting μ_1, μ_2 to be closer to each other and reducing variance. This approach generates perturbations which elicit more nuanced comparisons from annotators, allowing us to capture fine-grained legibility preferences in the dataset.

Inter-annotator Agreement. Three waves of annotations were collected using adaptive pair generation. To establish high quality and confidence in the test set labels, three annotations were collected for each pair of perturbations in the test set. Pairs where all annotators disagreed were removed from the test set. For 49.1% of pairs, all 3 annotators agree on the same label, 43.6% of pairs have agreement between 2 out of the 3 annotators, and only 7.3% of pairs have no agreement among annotators. Hence, even with 4 labels to choose from, for 92.7% of (w_1, w_2) pairs, at least two out of three annotators chose the same label. This suggests that the task is well-defined and has low variance.

	# pairs (w_1, w_2)	# distinct (w)	classification examples	ranking examples
Train	14622	4940	20217	9027
Val	3326	1140	4639	2013
Test	3712	1520	4774	2650
Total	21660	7600	29630	13690

Table 1: **LEGIT** dataset statistics. For each word, there exist multiple perturbed pairs, generated through three rounds of adaptive annotations.

Annotator Details. We recruit 150 annotators, all of whom had over a 95% acceptance rate for previous work done on the platform, as well as a history of over 1,000 completed tasks.⁵ Annotators are given occasional quality checks, wherein they annotate pairs drawn from a gold dataset labeled by the authors; annotators with less than 70% accuracy on the gold data were removed from the study and their annotations discarded from the final dataset. Annotators are given batches of 20 (w_1, w_2) pairs at a time; typically taking between 30 – 45 seconds to annotate. The average compensation per batch is \$0.12. Further details of the annotation interface and instructions are available in [Appendix A](#).

Hard Subsets. We identify challenging subsets of the collected data for the ranking and classification tasks. For ranking, the chosen subset ($N = 1052$) contains pairs (w_1, w_2) where $\frac{(n_1 - n_2)^2}{n_1 n_2} < 0.1$, i.e., both n ’s are close to each other, so it is hard in the sense that the perturbations have similar parameters ϕ but varying degrees of legibility—they cannot be ranked just by comparing metadata. For the classification task, the chosen subset ($N = 2626$) consists of all perturbations w_i with $n_i > 0.4$, making the task more challenging by excluding lightly-perturbed words which are easier to classify.

4 Tasks and Models

In this section, we start by introducing two tasks for characterizing the legibility of perturbed texts, followed by a number of models for solving them.

4.1 Tasks

From the labels collected in the previous section, we derive data for the two tasks: ranking and classification. The tasks assume that the original word

⁵686 annotators were excluded due to failing their first quality check. Many attempts were observed to be spam.

w is known, as we base our setup considering an attacker who is trying to find the best perturbation.

Ranking Task. Given a pair (w_1, w_2) of perturbations and the original word w as input, rank the perturbations in order of legibility. For this task, we only consider the subset of data labeled with strict rankings—i.e., excluding pairs labeled BL (Both Legible) and NL (Neither Legible). As the data is balanced, we only report accuracy as the main metric for this task.

Classification Task. Given a single perturbation w_i and the original word w , decide whether the perturbation is legible. While annotators performed pairwise comparison between (w_1, w_2), we can infer the binary legibility labels for w_i from pairwise rankings as follows: for labels BL and NL, we can make the obvious inference of *legible* and *illegible* for both w_i . For labels L_i , we can again infer that w_i is legible, but cannot say anything about $w_{j \neq i}$; all such w_j with unknown legibility are excluded from the classification task dataset. Since there are more legible than illegible instances in the data, we report both accuracy and F1 scores on this task.

4.2 Baselines

Majority Class. This baseline always predicts the majority class from the training set for every test example. For the ranking task, it always predicts w_2 as the preferred perturbation (resulting in an accuracy of 0.5), and for classification, the majority class is ‘legible’ (yielding 0.677 accuracy).

Logistic Regression using ϕ . Note that in an attack setting, the attacker would know the perturbation parameters ϕ exactly and may be interested in predicting the legibility of their perturbation using these parameters. Hence, we perform logistic regression directly on the attack parameters (n, k) to predict the label. Being a simple metadata-only baseline, this model does not take into account the characters that were perturbed or their position.

4.3 Text-based Models

ByT5. Legibility, as defined in this paper, is a *visual* property. However, we might expect pretrained language representations (e.g., those learned by large-scale language models) to also encode visual similarity between characters since the web-corpora used for pretraining might include similar-looking characters in the same contexts

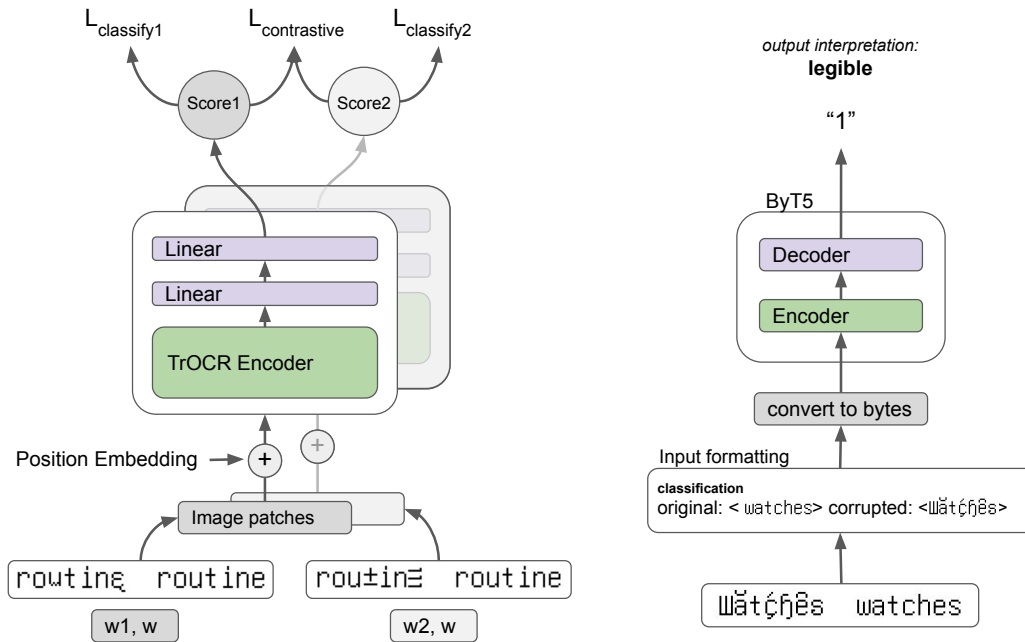


Figure 2: Comparing ByT5 and TrOCR training setup. ByT5: Both the perturbed and original words are given as one input to the model. TROCR: Both w_1, w_2 are fed sequentially into the same TrOCR-based model, and the two resulting scalar outputs are used to compute the loss. For each perturbation, the string “ $w_i w$ ” is rendered and used as input for the model.

(e.g., ‘0’ instead of ‘O’). To test this, we experiment with ByT5 (Xue et al., 2022), a multilingual encoder-decoder language model which tokenizes inputs into byte sequences. Byte-level tokenization ensures that none of the perturbations in **LEGIT** are out-of-vocabulary, and multilingual pretraining ensures that the model has seen a large subset of Unicode. We finetune the pretrained ByT5-models (‘small’ and ‘base’) to predict the binary labels for both classification and ranking in a text-in text-out setting. For ranking, the inputs are formatted as: “original: < w > word0: < w_0 > word1: < w_1 >”, and the output is “0” or “1” depending on which word is more legible. For classification, the inputs are formatted as: “original: < w > corrupted: < w_i >”, and the output is “0” or “1” depending on whether the corruption is illegible or legible. We train two separate models starting from the pretrained ByT5 weights using the cross-entropy loss over the target byte-sequence and AdamW optimizer (Loshchilov and Hutter, 2019) and perform early stopping using the validation set. Figure 2 outlines the model schematic with sample inputs and outputs.

4.4 Vision-based Models

Since we are concerned with finding representation spaces for *visually* similar characters, vision-based

models are a natural choice for the task. We consider both unsupervised models which rely on pixel-based or embedding-based similarities, as well as supervised models based on OCR, which we train on the **LEGIT** data.

IMGDOT. This unsupervised approach compares the corresponding characters in w and w_i based on the cosine distance between their pixel renderings. For the ranking task, this model selects the perturbation whose average cosine distance with the uncorrupted word is lower. For classification, we tune a threshold similarity parameter on the training set, above which the model predicts ‘legible’.

TROCR-Embeddings. This approach is identical to IMGDOT, except that we use the pretrained character embeddings obtained by passing the rendered images as input to the TROCR model. The embedding vector for each character is obtained by averaging the last hidden state from the encoder output (bypassing the pooler). Note that the accuracy of these unsupervised baselines gives us an idea of how well the corresponding representations align with human notions of legibility.

TROCR. Finally, we consider finetuning TROCR on the **LEGIT** data. We only use the encoder part of the TROCR base model and

connect it to a linear head. This linear head has two fully connected layers mapping inputs of size 768 (which is equal to the dimension of the encoder output) to a scalar output which represents the legibility score of the perturbed input. We use ReLU activations between the linear layers and apply dropout. The model takes variable-sized images as input; this is created by rendering a pair (w_i, w) into a single image by concatenating both strings along the horizontal axis (see Figure 2).

For the classification task, the output score from the model is used directly for predicting the label. Given a pair (w, w_i) , let s_i denote the scalar output from the model and let $y_i \in \{0, 1\}$ denote the legibility label (where 1 denotes that w_i is legible). Then the classification loss is given by:

$$\mathcal{L}_{\text{classify-}i} = -y_i \log \sigma(s_i) - (1 - y_i) \log [1 - \sigma(s_i)] \quad (1)$$

where σ is the sigmoid function. We apply the same loss function to both perturbations w_1 and w_2 . We denote this classification model as TROCR-C.

For the ranking task, we use the same model but apply it separately to the pairs (w, w_1) and (w, w_2) to obtain the scores s_1 and s_2 . The parameters across the two applications of the model are shared in a Siamese network setup (Koch et al., 2015). Given these two scores, and the label $y \in \{0, 1\}$ (where 0 denotes that w_1 is more legible), we define the ranking loss as:

$$\mathcal{L}_{\text{contrastive}} = -y \log \sigma(s_1 - s_2) - (1 - y) \log [1 - \sigma(s_1 - s_2)] \quad (2)$$

The above loss encourages s_1 to be higher than s_2 when $y = 0$ and vice versa. A similar loss has been used to train summarization models from pairwise human preferences (Stiennon et al., 2020). We denote this ranking model as TROCR-R.

The Siamese setup for the ranking task is limited in the sense that it cannot directly compare the two perturbations to decide which is more legible. However, our goal is to train the model to produce a calibrated legibility score given only a single perturbation as the input. Further, the Siamese network allows us to train the model on *both* the classification and ranking tasks together in a multi-task fashion:

$$\mathcal{L} = \mathcal{L}_{\text{classify-1}} + \mathcal{L}_{\text{classify-2}} + \mathcal{L}_{\text{contrastive}} \quad (3)$$

The loss terms for each training example are masked based on the label: the ranking loss is masked out if the label is “equally legible” or “both

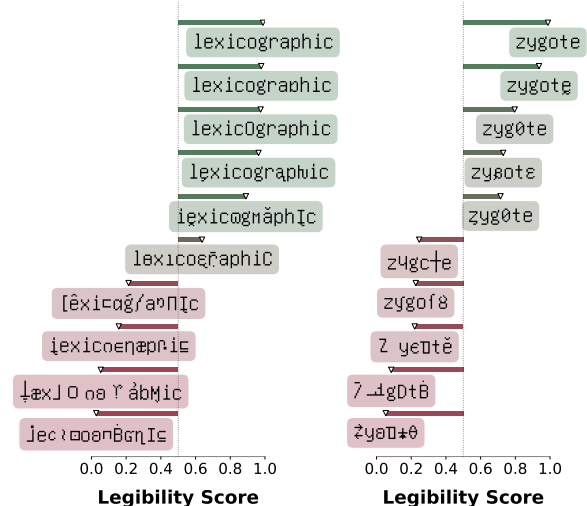


Figure 3: Legibility scores for LEGIT-generated perturbations of *lexicographic* and *zygote* from the TROCR-MT model. Neither word was seen during training.

unclear”, whereas the individual classify- i loss is masked out if the inferred binary legibility of perturbation w_i is indeterminate (e.g. for label L_1 , binary legibility of w_2 is unknown). Together, these losses ensure that the legibility score s_i is thresholded at 0, above which the perturbations are legible, and more legible inputs receive a higher score. We denote the model using combined loss as TROCR-MT.

5 Results

Table 2 shows the performance of all models introduced on both the classification and ranking tasks.

Classification Task. For the classification task, we find that baselines that just use the metadata perform poorly. The Majority Class baseline obtains an F1 score of 0.677, and the Logistic Regression model using ϕ parameters yields an F1 score of 0.665, implying that legibility is *not* a simple function of the perturbation parameters k, n . The unsupervised vision-based models, IMGDOT and TROCR embeddings, vastly improve upon the simple baselines, with the TROCR embeddings obtaining an F1 score 0.868 and IMGDOT yielding an F1 score of 0.845. Hence, these embeddings align reasonably well with human perceptions of legibility. The text-based ByT5 models improve significantly over the baselines and unsupervised vision-based models. They are comparable to the performance of the single-task objective TROCR-C, but worse than the TROCR-MT. This suggests that the ByT5 models might have encountered some visual perturbations during pretraining. Comparing the

Model		Classification				Ranking	
		Accuracy	Precision	Recall	F1 Std/Hard	Accuracy Std/Hard	
Baselines	Majority Class	.512	.512	1.000	.677/.000	.500/.502	
	Log. Regression	.680	.659	.671	.665/.256	.744/.642	
Vision-based	IMGDOT	.788	.861	.828	.845/.583	.790/.652	
	TrOCR embeds	.825	.868	.883	.868/.654	.781/.677	
	TrOCR-C	.840	.881	.891	.886/ -	-	
	TrOCR-R	-	-	-	-	.835/ -	
	TrOCR-MT	.868	.914	.895	.905/.726	.858/.757	
Text-based	ByT5-small	.844	.872	.909	.890/ -	.762/ -	
	ByT5-base	.842	.868	.912	.889/ -	.769/ -	

Table 2: Results on the standard test set. The TrOCR-MT model, trained in the multi-task setting, outperforms all other models for F1 score on both tasks. The trained models also outperform the baselines on both tasks.

single-task TROCR-C model with the multi-task TROCR-MT, we find that the presence of the additional ranking loss term during training improves model performance on the classification task from 0.886 to 0.905. On test examples where all 3 annotators agree, TROCR performs even better, attaining an F1 score of 0.960, compared to a score of 0.850 on examples where only 2 annotators agree. As further evidence of the model’s alignment with annotators, we find that the model confidence is directly correlated with annotator agreement (cf. Figure 4) as measured by Fleiss’ κ (Fleiss, 1971). Furthermore, consider Figure 3, which shows legibility scores obtained from TROCR-MT for two words picked at random which are not part of the training set. Qualitatively, we see that legibility scores from the TROCR-MT model aligns with the human judgements of legibility for these words.

Ranking Task. The TROCR-MT model performs better *relative* to other models, resulting in a 6.8% absolute accuracy improvement. Akin to the classification task, we find the TROCR-MT model outperforms its single-task counterpart TROCR-R. Thus, training with a multi-task objective improves performance on both ranking and classification tasks when compared to single-objective models. Differently from classification, we find that ByT5 is significantly worse than the vision-based models on ranking, suggesting that language model pretraining is effective at separating legible from illegible perturbations, but not at encoding the degree of legibility of legible perturbations.

Jigsaw Challenge. Next, we check whether perturbations generated by our attack model (§ 3.1) and filtered to ensure legibility using TROCR-MT are effective at degrading the performance

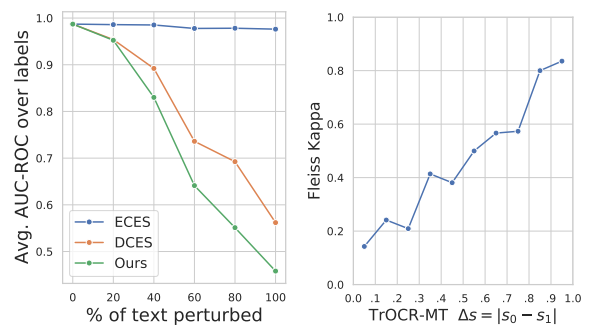


Figure 4: **Left:** *Detoxify* model performance on perturbations generated by different attack methods on a random subset ($N = 2000$) of the Jigsaw Toxic Comment Classification dataset. Model performance degrades most on our perturbations. **Right:** Model confidence on legibility is aligned with annotator agreement. Legibility scores (s_0, s_1) were obtained using TROCR-MT for each perturbed pair (w_0, w_1) in the test set. Pairs were grouped by the score difference $\Delta s = |s_0 - s_1|$ and Fleiss’ κ was computed for each group.

of NLP models. We employ the Jigsaw Toxic Comment Classification Dataset, which is a multilabel classification dataset consisting of Wikipedia comments and human-annotated binary labels for 6 toxicity categories. In Figure 4 (Left), we compare **LEGIT** and **VIPER-DCES** strategies in a real-world scenario by perturbing the Jigsaw dataset with each strategy and reporting how much these perturbations degrade the performance of *Detoxify*-original (Hanu and Unitary team, 2020), a BERT-based model which has state-of-the-art performance on the Jigsaw dataset. We show that **LEGIT** produces greater degradation at lower n , and produces more legible perturbations even at higher n (due to TROCR-MT filtering). In comparison, we find that **DCES** perturbations become very hard to read at higher n , diluting the signifi-

cance of the DCES results at high n . [Appendix D](#) provides a qualitative analysis of the legibility of DCES perturbations compared to those generated using our **LEGIT** method.

VIPER-ECES causes a negligible degradation on model performance, which is due to the fact that the BERT tokenizer “corrects” almost all of the simple diacritic-based ECES character substitutions. This means that the classification model receives mostly unperturbed input save for some isolated UNKs. For example, the perturbed input `Ťhãñķ ýôů` is tokenized back into `Thank you`. Taken together, these results demonstrate that **LEGIT** exploits a more *efficient* legibility space, finding character substitutions which have a greater impact on model performance while preserving legibility.

Perturbing GPT-3. The strong performance of ByT5 at separating legible from illegible inputs suggests that language models might be somewhat robust to such perturbations. To examine this, we experiment with GPT-3 (text-davinci-002 checkpoint) ([Brown et al., 2020](#)) using a *perturbation recovery task*, wherein we prompt the model to decode perturbed words back to their original strings. We sample a subset of 1,000 (w, w_i) pairs from **LEGIT** which have a label of *legible*. These perturbations are fed to the GPT-3 model in batches of 10, along with an instructional prompt (see [Appendix E](#)) and 4 examples; recovered words are received as a completion to the input prompt. In addition, we also perturb the same 1000 words using VIPER-DCES and report the accuracy of GPT-3 at reconstructing them. We observe that GPT-3 often returns a word with a short edit distance to the original word, and hence to capture this in our evaluation, we apply the Porter stemmer from NLTK ([Loper and Bird, 2002](#)) to both the original words and predicted reconstructions, and then measure how often their stemmed forms are the same. We repeat this experiment 3 times, randomly sampling the 4 examples in the prompt each time. [Figure 5](#) shows the GPT-3 accuracy at different fractions of corrupted characters ($n = \{0.3, 0.7, 1.0\}$). As expected, the accuracy goes down as n increases, but we find that GPT-3 performs worse on **LEGIT** perturbations. This demonstrates that while state-of-the-art language models are mildly robust to the narrower range of perturbations considered in existing visual attacks, they degrade significantly on inputs sampled from **LEGIT** which are marked by humans as legible. This result underscores the im-

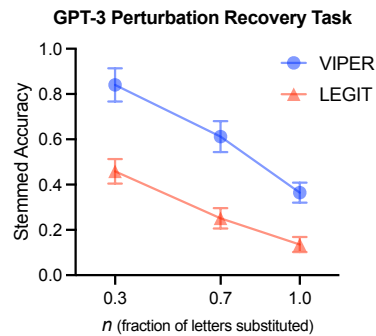


Figure 5: **LEGIT** perturbations sampled at low n degrade accuracy at levels comparable to the $n = 1$ VIPER configuration. Error bars indicate 95% confidence interval.

portance of considering the entire space of legible perturbations when evaluating model robustness.

6 Conclusion

We set out to characterize the limits of legibility of visual perturbations. To do so, we first collected and released a new dataset, **LEGIT**, comprising legibility preferences of human subjects. Using this dataset, we framed a binary legible-or-not classification task, and a ranking task to rank candidate perturbations. For these tasks, we explored several text- and vision-based models, and found that our models obtain a high F1 score of 0.91 for the classification task and an accuracy of 0.86 for the ranking task. Perturbations generated using the same attack method as used for constructing **LEGIT** lead to significant degradation on the Jigsaw Challenge task and are not recovered by GPT-3 accurately, despite being filtered for legibility. We believe this work opens avenues for research on legibility-driven certified robustness to visual attacks in NLP.

Limitations

At the outset, we note that while our legibility-scoring models are a step forward towards defending against visual attacks, they should not be seen as perfect. Defending against all of the attacks which our models find legible might still leave room for legible attacks missed by our system.

Moreover, we note that the perturbation procedure outlined here only generates substitution-based perturbations. Whereas, characters may also be deleted, added, or swapped, and multiple adjacent characters may be substituted with visually similar counterparts (see [Figure 1](#)). Future work may explore broader classes of perturbations.

When constructing the dataset, we only chose words with a length of at least 4 letters, excluding many common 3-letter words. This is because for 3-letter words, there is a high likelihood that a bad perturbation may be mistakenly recognized as a good perturbation by virtue of being in-vocabulary. For example, “ban” is a bad perturbation of “man”, but for an annotator who sees it without knowing the original word and in absence of any sentence-level context, it seems like a perfectly good perturbation, when in fact it obscures the meaning of the original word. This is a limitation of the experimental setup that can lead to bad annotations, and to mitigate it we chose a higher minimum word length as longer words have fewer such collisions.

Further, we study word-level perturbations in isolation without any surrounding context, whereas in practice, readers often can decipher words based on the context. In general, the legibility of a text depends on the context around it—for example, even if a word is deleted from a sentence it is often possible to reconstruct it. The data we collect here, however, measures the legibility of individual words without any context, in order to simplify the generation and annotation process. As a result, the legibility estimated using this data should be considered as *lower bound* of the legibility in any given context. This was a deliberate choice as we wanted to ensure that whatever we ascertain as legible is legible in *all* contexts.

Lastly, the models we develop in our work are of relatively moderate size (334 – 584 million parameters) and take only unimodal input (i.e. pixels for TROCR models and Unicode bytes for ByT5). and future work may be able to improve the performance by using larger models which accept multimodal input (e.g. both pixels and Unicode bytes simultaneously) and learn joint representations across these modalities.

Ethical Considerations

The word list comprising our dataset was filtered to remove swear words, slurs etc. in order to avoid exposing annotators to potentially harmful content.

Acknowledgements

We thank Professors Carlo Tomasi and Sam Wiseman at Duke University for their helpful feedback. This research was supported by a grant from the Arts & Sciences Council Committee on Faculty Research at Duke University.

References

- Moustafa Alzantot, Yash Sharma, Ahmed Elgohary, Bo-Jhang Ho, Mani Srivastava, and Kai-Wei Chang. 2018. [Generating natural language adversarial examples](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2890–2896. Association for Computational Linguistics.
- Yonatan Belinkov and Yonatan Bisk. 2018. [Synthetic and natural noise both break neural machine translation](#). In *International Conference on Learning Representations*.
- Nicholas Boucher, Iliia Shumailov, Ross Anderson, and Nicolas Papernot. 2022. Bad characters: Imperceptible nlp attacks. In *2022 IEEE Symposium on Security and Privacy (SP)*, pages 1987–2004. IEEE.
- Thorsten Brants and Alex Franz. 2006. Web 1t 5-gram ver. 1. *LDC2006T13, Linguistic Data Consortium, Philadelphia*.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. [Language models are few-shot learners](#). In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.
- Chris Callison-Burch, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. 2007. [\(meta-\) evaluation of machine translation](#). In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 136–158, Prague, Czech Republic. Association for Computational Linguistics.
- Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. 2020. End-to-end object detection with transformers. In *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*, pages 213–229. Springer.
- Alex Clark. 2015. [Pillow \(pil fork\) documentation](#).
- Mark Davis and Michel Suignard. 2021. [UTS #39: Unicode Security Mechanisms. Version 14.0](#).
- Antreas Dionysiou and Elias Athanasopoulos. 2021. [Unicode Evil: Evading NLP Systems Using Visual Similarities of Text Characters](#). In *Proceedings of the 14th ACM Workshop on Artificial Intelligence and Security, AISec ’21*, pages 1–12, New York, NY, USA. Association for Computing Machinery.

- Javid Ebrahimi, Daniel Lowd, and Dejing Dou. 2018. [On Adversarial Examples for Character-Level Neural Machine Translation](#). In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 653–663, Santa Fe, New Mexico, USA. Association for Computational Linguistics.
- Steffen Eger and Yannik Benz. 2020. [From Hero to Zéro: 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*, pages 786–803, Suzhou, China. Association for Computational Linguistics.
- Steffen Eger, Gözde Gül Şahin, Andreas Rücklé, Ji-Ung Lee, Claudia Schulz, Mohsen Mesgar, Krishnkant Swarnkar, Edwin Simpson, and Iryna Gurevych. 2019. Text processing like humans do: Visually attacking and shielding nlp systems. 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)*, pages 1634–1647.
- Eveline Flamand. 2008. Deciphering 133t5p34k internet slang on message boards. *MS thesis*.
- Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.
- Siddhant Garg and Goutham Ramakrishnan. 2020. Bae: Bert-based adversarial examples for text classification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6174–6181.
- Laura Hanu and Unitary team. 2020. Detoxify. Github. <https://github.com/unitaryai/detoxify>.
- Robin Jia and Percy Liang. 2017. Adversarial examples for evaluating reading comprehension systems. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2021–2031.
- Di Jin, Zhijing Jin, Joey Tianyi Zhou, and Peter Szolovits. 2020. Is bert really robust? a strong baseline for natural language attack on text classification and entailment. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 8018–8025.
- Josh Kaufman. 2012. google-10000-english. <https://github.com/first20hours/google-10000-english>.
- Gregory Koch, Richard Zemel, Ruslan Salakhutdinov, et al. 2015. Siamese neural networks for one-shot image recognition. In *ICML deep learning workshop*, volume 2, page 0. Lille.
- Thai Le, Jooyoung Lee, Kevin Yen, Yifan Hu, and Dongwon Lee. 2022. [Perturbations in the Wild: Leveraging Human-Written Text Perturbations for Realistic Adversarial Attack and Defense](#). In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2953–2965, Dublin, Ireland. Association for Computational Linguistics.
- Jinfeng Li, Shouling Ji, Tianyu Du, Bo Li, and Ting Wang. 2019. [TextBugger: Generating Adversarial Text Against Real-world Applications](#). *Proceedings 2019 Network and Distributed System Security Symposium*.
- Linyang Li, Ruotian Ma, Qipeng Guo, Xiangyang Xue, and Xipeng Qiu. 2020. Bert-attack: Adversarial attack against bert using bert. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6193–6202.
- Minghao Li, Tengchao Lv, Lei Cui, Yijuan Lu, Dinei Florencio, Cha Zhang, Zhoujun Li, and Furu Wei. 2021. [TrOCR: Transformer-based Optical Character Recognition with Pre-trained Models](#). Technical Report arXiv:2109.10282, arXiv. ArXiv:2109.10282 [cs] type: article.
- Weixin Liang, James Zou, and Zhou Yu. 2020. Beyond user self-reported likert scale ratings: A comparison model for automatic dialog evaluation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1363–1374.
- Edward Loper and Steven Bird. 2002. [Nltk: The natural language toolkit](#). *CoRR*, cs.CL/0205028.
- Ilya Loshchilov and Frank Hutter. 2019. [Decoupled weight decay regularization](#). In *International Conference on Learning Representations*.
- John X. Morris, Eli Lifland, Jin Yong Yoo, Jake Grigsby, Di Jin, and Yanjun Qi. 2020. [TextAttack: A framework for adversarial attacks, data augmentation, and adversarial training in NLP](#). Version: 4.
- Danish Pruthi, Bhuwan Dhingra, and Zachary C Lipton. 2019. Combating adversarial misspellings with robust word recognition. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5582–5591.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.
- Eric S Raymond. 1996. *The new hacker’s dictionary*. Mit Press.
- Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. 2020. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, 33:3008–3021.

Linting Xue, Aditya Barua, Noah Constant, Rami Al-Rfou, Sharan Narang, Mihir Kale, Adam Roberts, and Colin Raffel. 2022. ByT5: Towards a token-free future with pre-trained byte-to-byte models. *Transactions of the Association for Computational Linguistics*, 10:291–306.

A mTurk Annotation Interface

The web-based UI used by mTurk Annotators is shown in Figure 6, with the instructions being visible throughout the duration of the task. Note that the perturbed words w_1, w_2 are rendered in GNU Unifont, which is the same font that words are rendered in for computing visual similarity (cf. §Legibility Tests, Perturbation Process). This ensures that both annotators and visual similarity models see pixel-for-pixel identical perturbations, controlling for the fact that different fonts render the same character differently.

The interface is optimized for clarity and labeling speed, with a focus on eliminating unnecessary UI elements and minimizing clicks. Labeling each pair w_1, w_2 with one label $L \in \{L_1, L_2, BL, NL\}$ takes exactly one click. Annotators choose L_1 by clicking on w_1 (the left word), and similarly by clicking on w_2 (the right word) for w_2 . BL is selected using the “equally legible” button, whereas NL is chosen by clicking on “both unclear.”

Immediately after a choice is made, the UI updates and the next pair in the batch is shown (there is no option to go back and edit the chosen label). Annotators who attempt to cheat on the task by “speeding through” (i.e. clicking randomly or spamming the same choice) end up failing the occasionally administered quality checks and are subsequently disinvited from the study.

B Use of OCR Models

Boucher et al. (2022) propose using Optical Character Recognition (OCR) models to preprocess input for text-based language models. Rendering input text and passing it through an OCR before giving it to the language model filters certain kinds of misleading Unicode characters (e.g. invisible control sequences or near-identical Confusables (Davis and Suignard, 2021)) from the text. However, when used for legible but visually distinct perturbations, off-the-shelf OCR models run into two problems.

Firstly, both mono- and multi-lingual OCR models will recognize characters from learned scripts at face value, instead of recognizing their intended use as visually similar substitutions. For example,

TROCR (Li et al., 2021), when given an image of the string ‘Mex!(0’, decodes it into ‘Mex!(0’ (i.e. the same string), completely ignoring its intended meaning (Mexico). Secondly, since OCR models are only trained on semantically meaningful inputs, they do not learn good priors to differentiate non-sense inputs from highly perturbed inputs.

We use two OCR-capable models on the ranking and classification tasks: TROCR, which is explicitly trained on an OCR dataset, as well as CLIP, which is trained on a general corpus containing images of texts from which it learns “a high quality semantic OCR representation that performs well on digitally rendered text” (Radford et al., 2021).

We find that TROCR models fine-tuned on our dataset achieve high performance on legibility-related tasks. On the text side, we consider the token-free language model, ByT5 (Xue et al., 2022), which encodes each byte individually, as opposed to byte-pairs or subword tokens longer than one byte. Since its encoding of each byte is disentangled from surrounding bytes, ByT5 is able to retain a larger share of the unperturbed part of the string, hopefully making it more robust to character-substitution perturbations compared to token-based models, which reduce the sequences with perturbed characters into rare tokens or simply to UNKS.

C Hyperparameters

TROCR (‘base-handwritten’ version) was fine-tuned on LEGIT with the loss function configurations (C, R, MT) described above. To train each configuration, we use a single NVIDIA A6000 GPU (48GB VRAM) with a batch size of 26 and learning rate of 10^{-5} with the AdamW optimizer and a linear decay schedule (without warmup). ByT5-base and ByT5-small were trained on the same hardware with a batch size of 8 and learning rate of 10^{-4} .

D Toxic Comment Classification Experiment

The original string from the Jigsaw Toxic Comment Classification dataset is:

It is needed in this case to clarify that UB is a SUNY Center. It says it even in Binghamton University at Albany, State University of New York, and

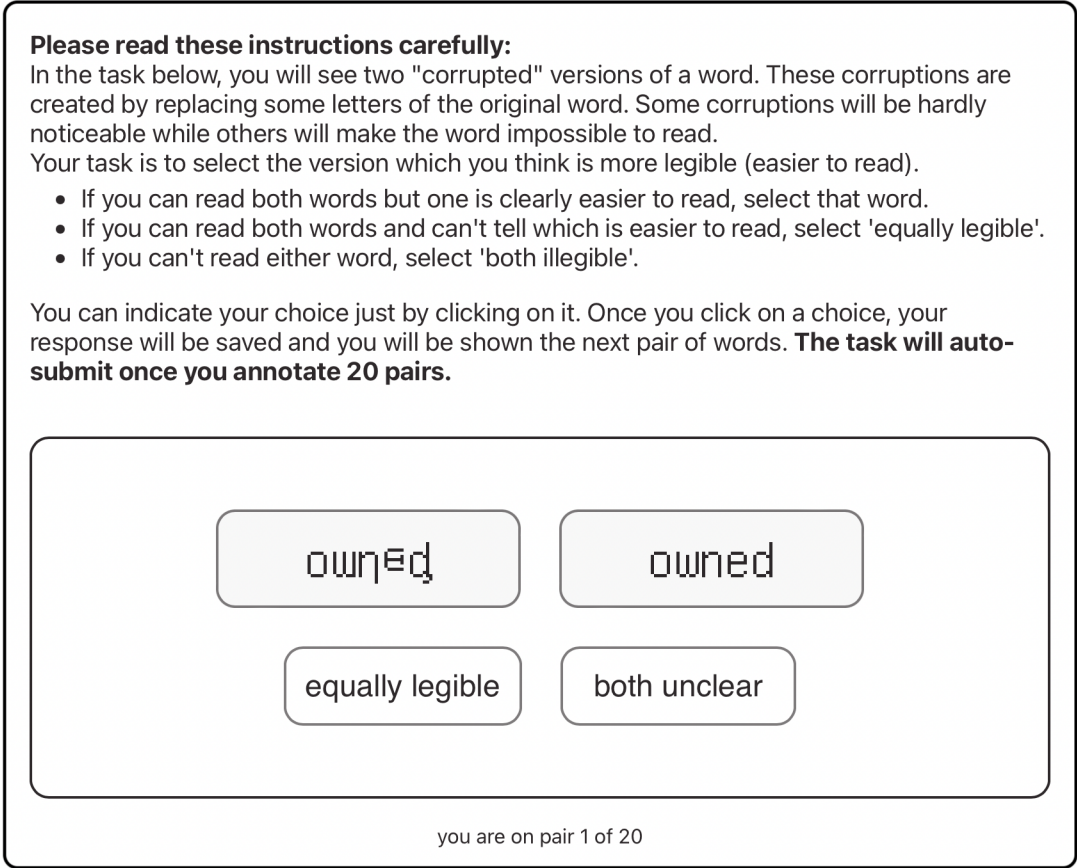


Figure 6: The mTurk Annotation Interface

Stony Brook University. Stop trying to say it's not because I am totally right in this case.

1. o_1
2. o_2
3. o_3
4. o_4
- 5.

The VIPER DCES and **LEGIT** perturbations are compared in Figure 7. The **LEGIT** perturbations were labeled as legible by TROCR-MT.

E GPT-3 Experiment

We provided the following prompt to the text-davinci-002 checkpoint using the GPT-3 API:

The model is allowed to condition on 4 ground-truth examples: o_1 through o_4 , and attempts to generate o_5 through o_{10} by providing a completion for the prompt above. The *temperature* and *top p* parameters were both set to 1 to allow for consistent and reproducible outputs across batches.

The following is a list of corrupted words and their correct versions. The corruptions were created by replacing some or all letters of the correct version with similar-looking letters.

- Corrupted:
1. c_1
 2. c_2
 - ...
 10. c_{10}
- Original:

It is needed in this case to clarify that UB is a SUNV Center. It was in the Birmingham University, University at Alton, State University of New York and Stony Brook University. Stop trying to be busy in am total in this case.

(a) VIPER DCES, $n = 1.0$, nearest neighbors sampled uniformly from list of top 10 neighbors for each character.

It is needed in this case to clarify that UB is a SUNV Center. It was in the Birmingham University, University at Alton, State University of New York and Stony Brook University. Stop trying to be busy in am total in this case.

(b) Ours (LEGIT perturbation strategy with TROCR-MT legibility filter) $n = 1.0$, nearest neighbors sampled normally ($\mu = 15$, $\sigma^2 = 7$) from top 30 neighbors for each character

Figure 7: A randomly selected paragraph from the Jigsaw dataset (a) perturbed by VIPER DCES (b) and our method (c). Our perturbation appears more legible despite being generated using harsher parameters.