Some Odd Adversarial Perturbations and the Notion of Adversarial Closeness

Shakila Mahjabin Tonni[1*] and **Pedro Faustini**[2] and **Mark Dras**[2]
CSIRO's Data61, Sydney, Australia[1]
School of Computing, Macquarie University, Sydney, Australia[2]
*shakila.tonni@csiro.au**

Abstract

Deep learning models for language are vulnerable to adversarial examples. However, the perturbations introduced can sometimes seem odd or very noticeable to humans, which can make them less effective, a notion captured in some recent investigations as a property of '(non-)suspicion'. In this paper, we focus on three main types of perturbations that may raise suspicion: changes to named entities, inconsistent morphological inflections, and the use of non-English words. We define a notion of adversarial closeness and collect human annotations to construct two new datasets. We then use these datasets to investigate whether these kinds of perturbations have a disproportionate effect on human judgements. Following that, we propose new constraints to include in a constraint-based optimisation approach to adversarial text generation. Our human evaluation shows that these do improve the process by preventing the generation of especially odd or marked texts.

1 Introduction

Adversarial examples, aimed at deceiving machine learning models by making subtle changes to the inputs, are often highly successful (Li et al., 2019; Eger and Benz, 2020; Garg and Ramakrishnan, 2020, for example). Three such adversarial text examples are given in Table 1. Several adversarial attacks on Large Language Models (LLMs) also show success, where the attacker prompts LLMs to generate context-preserving word replacements (Wang et al., 2024) or manipulates the LLMs through prompt injection (Shi et al., 2022; Zou et al., 2023). Relatedly, there have been a few attempts to humanise LLM-generated sentences by applying adversarial attacks on them to bypass AI-text detectors (Zhou et al., 2024; Cheng et al., 2025). For image (Szegedy et al., 2014; Carlini and Wagner, 2017; Ma et al., 2018, for example), the added noise in the input is typically required to be imperceptible to humans, such that if the original and adversarial variants were side by side, the change would not be noticeable. However, texts are more complex as the modifications in this context are visible, and they have to maintain semantic similarity to the original.

In scenarios where human perception matters—such as reviews attempting to bypass filters that are intended to be read by humans, or phishing emails—adversarial texts should not be dismissed by readers as machine-generated. To this end, Morris et al. (2020a) introduced the concept of *(non-)suspicion*, which focuses on whether a human reader can detect that the text has been modified. This idea, along with other constraints like semantics and grammaticality, provides a new lens to evaluate adversarial examples in NLP. Dyrmishi et al. (2023) expanded this to a more comprehensive analysis across several adversarial attacks, datasets, and attributes of generated adversarial texts. Tonni et al. (2025) followed that by considering suspiciousness levels as graded rather than binary and used suspicion scores predicted by a regression model to successfully generate less suspicious-looking sentences.

In this work, we consider that particular types of adversarial perturbations might strike human readers as odd, such that they should be avoided if the goal is to generate more natural-looking adversarial texts. Specifically, we examine three such perturbations: changes to named entities (NEs), changes to morphological inflections or parts of speech such that the results are inconsistent with the rest of the text, and changes that introduce words not in the original language. A few examples of all these problematic scenarios are illustrated in Table 1 on a few texts from a review website.

To determine whether, in fact, humans do perceive adversarial texts with these perturbations as particularly poor, we define a notion of *adversarial closeness* — a measure of the human perception

Proceedings of The 23rd Annual Workshop of the Australasian Language Technology Association, pages 28–44
November 26-28, 2025

orig. — every time I've ordered something from Playcom something has gone wrong and that covers the last five years stick to Amazon or eBay **adv.** — every time i've ordered something from playcom something has gone amiss and that **blanket** the iast five **aged** stick to amazon or ebay
orig. — having spent a lot of time and money importing gbics and dac cables from the us and china i decided to leave the hassle **adv.** — having spent a lot of time and money importing gbics and dac yarn from the us and **porcelain** i decided to leave the hassle
orig. — a saving of over is anticipated against my next full years bill **adv.** — a saving of over is anticipated against my next holistic aged **lois**

Table 1: Adversarial texts generated from the TRUSTPILOT dataset, with inflectional changes (top), named entity changes (middle) and foreign language changes (bottom). Changed words are in blue, with those related to the specified perturbations in **bold**. See Sec 4.

of how similar adversarial texts are to the originals. This goes beyond the notions of (non-)suspicion noted above, which focus on the suspiciousness of the adversarial sentence in isolation, whereas this judgement is made based on both an original text and its adversarial variant. The goal is to generate better adversarial variants (e.g., in privacy applications (Faustini et al., 2025)), where the generator will likewise have access to the original text, and can potentially take advantage of the extra information. While previous human evaluations have focused on aspects like grammaticality or semantic correctness (Garg and Ramakrishnan, 2020; Jin et al., 2020; Li et al., 2019) separately, they don't fully capture the overall human perception of adversarial sentences.

We then gather two datasets of human judgements of adversarial closeness and use these to analyse the effects on human judgements of our three chosen perturbations, additionally supplementing this with an analysis using automated readability metrics. Following this, we add constraints on these three perturbation types into an adversarial generation method to assess whether this can be used to improve human perceptions of adversarial texts. We then carry out an evaluation comparing regular adversarial texts and those with the additional constraints. All our data and codes are available.[1]

Overall, our contributions are three-fold:

1. Developing and publishing an annotated dataset of human perceptions of the "adversarial closeness" between human-written texts and their adversarial counterparts, complementary to human suspiciousness.
2. Assessing the impact of selected perturbations according to these adversarial closeness scores, as well as on an automatic reading ease metric.
3. Implementing new constraints on adversarial sentence generation based on these selected perturbations and carrying out an evaluation that demonstrates that these produce better quality adversarial sentences.

[1] https://github.com/SJabin/AdversarialCloseness

2 Related Works

Adversarial Text Evaluation. Morris et al. (2020a) outlined four core properties of adversarial text quality: semantic similarity, grammaticality, overlap with the original text, and non-suspicion of the human readers. These primary evaluation metrics of adversarial texts are typically incorporated by the algorithms as sentence-generation constraints (Jin et al., 2020). Semantic similarity to the original text is measured using metrics such as Universal Sentence Encoder (USE)-based cosine similarity (Jin et al., 2020; Garg and Ramakrishnan, 2020; Li et al., 2021; Liu et al., 2023; Chi et al., 2022), or using a fine-tuned model as in Yoo and Qi (2021). Grammaticality is assessed using the word recognition model in Pruthi et al. (2019) or part-of-speech-preserving word replacements in (Jin et al., 2020; Ebrahimi et al., 2018; Garg and Ramakrishnan, 2020). Overlap measures, such as Levenshtein edit distance (Gao et al., 2018) and n-gram-based metrics like BLEU (Wang et al., 2020; Yildiz and Tantuğ, 2019), quantify differences between original and adversarial texts.

Human Evaluation. Some adversarial algorithms rely solely on automatic metrics for evaluation, while others incorporate human assessments. Common human evaluation metrics include human classification accuracy for the original task (Jin et al., 2020; Alzantot et al., 2018; Garg and Ramakrishnan, 2020; Li et al., 2020), the similarity of adversarial texts to the original (Jin et al., 2020; Alzantot et al., 2018; Li et al., 2023, 2021, 2020), and grammatical correctness (Jin et al., 2020; Li et al., 2023, 2021, 2020). Beyond these, some studies also assess the naturalness of adversarial sentences as in BAE (Garg and Ramakrishnan, 2020) and detectability as in PWWS (Ren et al., 2019) by

having human readers compare original-adversarial sentence pairs.

Human Suspiciousness. The work of Morris et al. (2020a) introduced 'non-suspicion' as a crucial property to evaluate. On a movie review dataset applying TEXTFOOLER (Jin et al., 2020) and GENETICATTACK (Alzantot et al., 2018), they asked human judges to rate semantic preservation (1–5 Likert scale), grammaticality (identifying errors), and suspiciousness determined by whether judges identified sentences as real or computer-altered. They obtained 10 judgements per text and found that 69.2% of TEXTFOOLER examples were judged suspicious. Expanding this, Dyrmishi et al. (2023) conducted a large-scale survey across 9 word-based attacks and 3 datasets, evaluating validity, naturalness (comprising suspiciousness, detectability, grammaticality, and meaningfulness), and word-level detectability. They found that 60.33% of adversarial texts were judged as computer-altered, compared to 21.43% of original texts, and humans detected 45.28% of perturbed words in adversarial texts. Following a similar setting, Tonni et al. (2025) collected graded suspiciousness scores on a 1-5 Likert scale by human judges on a set of adversarial movie reviews generated by TEXTFOOLER. They also built regressors to predict human suspiciousness levels and adopted the regressors to produce less suspicious sentences.

3 Adversarial Closeness Datasets

We define adversarial closeness as a measure of how similar adversarial texts are to the original sentences from the perspective of human readers. In this section, we give details of preparing the adversarial closeness dataset.

3.1 Collecting Human Annotations

On Amazon Mechanical Turk (MTurk),[2] we present the original and adversarial sentences in pairs to human judges, who annotate the extent to which the adversarial texts remain close to the original texts on an ordinal Likert scale ranging from 1 (very close) to 5 (very different). We also marked the perturbed words in both the real and adversarial sentences with a distinct colour. MTurk survey details and sample interface are in App D.

Due to the paired sentence scenario, we limit to one adversarial attack method to avoid complications in scoring n-way comparisons (1 original and $n-1$ adversarial alternatives). We use TEXTFOOLER (Jin et al., 2020) for two reasons: (1) it performed well in the suspiciousness analyses of Tonni et al. (2025), and (2) it was a key method in the work of Morris et al. (2020a) and Dyrmishi et al. (2023). We prepare the following annotated datasets:

	MOVIEREVIEW		TRUSTPILOT	
Score	Freq.	%	Freq.	%
1	103	17.08	176	9.77
2	150	24.88	503	27.93
3	146	24.21	673	37.38
4	87	14.43	363	20.21
5	117	19.40	85	4.71
Avg score	2.94		2.82	

Table 2: Human score distribution of adversarial closeness for MOVIEREVIEW and TRUSTPILOT. Score range is 1 ("very close") to 5 ("very different").

TRUSTPILOT (Hovy et al., 2015). This dataset has customer reviews from the English subset ('en-uk') with $985,106$ train records and $364,855$ test records. It has, on average, 64 words and a maximum of 1136 words per sentence. We sampled $200,000$ training and $5,000$ testing records to fine-tune a pre-trained $\text{BERT}_{\text{BASE}}$ model, choosing "gender" to be the classification label with "male" and "female" classes.[3] We apply TEXTFOOLER on the 4,095 correctly predicted original sentences and generate 3,426 successful adversarial sentences. The test accuracy of the model is 0.819, which goes down to 0.088 under attack.

We then select 1800 sentences for our human evaluation and collect single judgements on 1500 sentence pairs (the DISTINCT SET) and collect 3 judgements on 300 sentence pairs (the COMMON SET), a total of 2400 annotated sentences.

MOVIEREVIEW (Pang and Lee, 2005). The Rotten Tomatoes Movie Review (MOVIEREVIEW) sentiment analysis dataset[4] is obtained from Tonni et al. (2025), with human suspiciousness annotation for 1206 original and TEXTFOOLER generated sentences (603 pairs). We further collect single annotations for closeness judgement on 540 sentence pairs (distinct set) and 3 annotations on 63 sentence pairs (common set).

[2] https://www.mturk.com/

[3] Binary categorisation is in line with the original setup of the dataset of Hovy et al. (2015).

[4] Orig. data source: https://tinyurl.com/mssr27tr

3.2 Analyses

Main results. The score distribution is presented in Table 2. Here, we use the median scores for the common set. The average closeness scores are 2.94 for MOVIEREVIEW and 2.82 for TRUSTPILOT, very close to the midpoint 3, suggesting the annotators have calibrated to the scale used.

Only 17.08% MOVIEREVIEW and 9.77% TRUSTPILOT sentences are scored as 1, suggesting that the adversarial perturbations in general are not very close to the real sentences. High percentages for the scores 4 (14.43%) and 5 (19.40%) imply higher divergence from the original sentences on MOVIEREVIEW, which is similar for TRUSTPILOT (20.21% scored 4 and 4.71% scored 5). These results indicate that the perturbations frequently fail to preserve key aspects of the original sentence.

MOVIEREVIEW and TRUSTPILOT score distributions are fairly similar, with MOVIEREVIEW scores being less concentrated around 3. Using Shannon entropy as a measure of distribution flatness, the score of 0.690 on MOVIEREVIEW indicates that the scores are slightly more evenly distributed than those on TRUSTPILOT, with a score of 0.616.

Differences between our *closeness* measure and the *suspiciousness* measure of Tonni et al. (2025) on MOVIEREVIEW dataset are in App. A .

In further analysis, we report the weighted average, taking the skewness of the data distribution into account.

Inter-annotator agreement. On the common set of MOVIEREVIEW and TRUSTPILOT sentences, we analyse the level of agreement among the three annotators. We note that although there is some degree of common practice in NLP regarding metrics for evaluating agreement on labels representing a *nominal* factor (e.g., Cohen's kappa), this is not the case for *ordinal* factors like ours. Following Tonni et al. (2025), we therefore use for our adversarial closeness scale the approach of Vogel et al. (2020), who calculate the average absolute deviation from the median response of the judges for each sentence; lower deviation means higher agreement. For C annotators and k categories of the Likert scale, the normalised annotator disagreement of each sentence i is:

$$\delta_i = \frac{\sum_{c=1}^{C} \left| \tilde{k}_i - k_i^c \right|}{C} \quad (1)$$

where $\tilde{k}_i$ is the median response to a sentence i and k_i^c is the category supplied by annotator c to i. Overall disagreement level δ is calculated by taking the average of the δ_i.

	MOVIEREVIEW		TRUSTPILOT	
Overall δ	0.61		0.62	
	Freq.	%	Freq.	%
$C\delta_i$=0	9	14.29	21	7.00
$C\delta_i$=1	18	28.57	97	32.33
$C\delta_i$=2	18	28.57	104	34.67
$C\delta_i$=3	11	17.46	63	21.00
$C\delta_i$=4	7	11.11	15	5.00
Wgt. avg.	1.846		1.851	

Table 3: Average inter-annotator disagreements δ and the frequency of δ_i levels on MOVIEREVIEW and TRUSTPILOT closeness evaluation.

The overall disagreement level δ and the unnormalised per-item disagreement frequency $C\delta_i$ for both the MOVIEREVIEW and TRUSTPILOT are presented in Table 3. δ is 0.61 for MOVIEREVIEW and 0.62 for TRUSTPILOT. To interpret these numbers: for our Likert scale of 5 points with $C = 3$ annotators, the smallest possible value for δ is 0 (perfect agreement) and the largest possible value is 1.33, giving a middling level of agreement. Surprisingly, for MOVIEREVIEW, even with the added knowledge of the original texts, we get the same level of disagreement of $\delta = 0.61$ as the suspiciousness ratings from Tonni et al. (2025) (where they also use a 5-point Likert scale, so operate on the same δ range), even though we might have expected the suspiciousness scores (with less context provided) to be more variable. Also, the perception levels have notable variations, with a very small amount of unanimous ($C\delta_i = 0$) judgements, 14.29% on MOVIEREVIEW and 7% on TRUSTPILOT.

4 Selected Perturbations

An analysis of the adversarial sentences from Tonni et al. (2025) reveals that highly suspicious sentences have certain perturbed elements, which likely contribute to their high suspiciousness scores. We observe the same pattern in the dataset from Sec. 3. In this section, we describe the three types of perturbed elements that we focus on, with examples.[5] We then use the TRUSTPILOT dataset to analyse the relationship of these perturbed elements to adversarial closeness scores, and also to automated readability scores.

[5]Examples are available in the appendix, from MOVIEREVIEW in Table 14 and from TRUSTPILOT in Table 15.

	Overall		INFLECTPERT		NEPERT		LANGPERT	
score	Freq.	%	Freq.	%	Freq.	%	Freq.	%
1	176	9.78	52	8.70	3	3.00	1	4.55
2	503	27.94	175	29.26	21	21.00	9	40.91
3	673	37.39	216	36.12	45	45.00	7	31.82
4	363	20.17	126	21.07	23	23.00	4	18.18
5	85	4.72	29	4.85	8	8.00	1	4.55
Total	1800		598		100		22	
Wgt. Avg.	2.82		2.84		3.12		2.78	

Table 4: Human judgement distribution for the three types of token perturbations on TRUSTPILOT, with scores from 1 ("very close") to 5 ("very different").

	Overall		INFLECTPERT		NEPERT		LANGPERT	
δ	0.62		0.62		0.72		0.73	
	Freq.	%	Freq.	%	Freq.	%	Freq.	%
$C\delta_i$=0	21	7.00	4	3.31	3	3.00	0	0.00
$C\delta_i$=1	97	32.33	45	37.19	23	23.00	1	16.67
$C\delta_i$=2	104	34.67	41	33.88	39	39.00	2	33.33
$C\delta_i$=3	63	21.00	27	22.31	25	25.00	2	33.33
$C\delta_i$=4	15	5.00	4	3.31	10	10.00	1	16.67
Total	300		121		100		6	
Wgt. avg.	1.85		1.85		2.16		2.5	

Table 5: Inter-annotator disagreement level of the TRUSTPILOT TEXTFOOLER sentences for INFLECTPERT, NEPERT and LANGPERT.

4.1 Perturbation Types

NEPERT. This type covers altering the named entities (NEs) in sentences. TRUSTPILOT in particular, due to being a product and service review dataset, has a diverse range of NEs, perturbing which may lead to unnatural sentences. In the 4th sentence from Table 15, the word "china" (which is clearly used in the sense of the country, as it is paired with the US) is replaced with "porcelain" in the below sentence with a closeness score of 5 from the human adversarial closeness data (very different).

INFLECTPERT. In this type of perturbation, the POS tag of a perturbation does not agree with the possible inflection POS tags of the lemmatised token from the original word. To illustrate, consider the 1st sentence from Table 15, with a closeness score of 3. In this case, the word "technical", is an adjective (POS: 'JJ') and a lemma with no other possible inflections. It is replaced by the word "technician" (noun -'NN'), which doesn't align with the possible inflected POS tags.

LANGPERT. We exclude from our analysis input sentences that are completely or partially written in another language.[6] Our interest is in the scenarios when the adversarial algorithm substitutes English words with words from different languages. For example, the 5th sentence in Table 15, the word "bill" was altered to "lois" (French for "laws") with closeness score of 4.

Counting up the occurrences of these perturbation types in the TRUSTPILOT dataset, we found that TEXTFOOLER produces 100 sentences with NEPERT, 598 INFLECTPERT and 22 LANGPERT.

[6]Often writers use non-English words or sentences to express strong opinions in reviews or posts.

4.2 Analysis

Connection with human annotation. From Table 4, among the 100 NEPERT sentences, 45 sentences are scored as 3, 23 as 4 and 8 as 5, totalling 76 sentences that are identified to be fairly-to-very different from the originals. Similarly, 371/598 INFLECTPERT sentences and 12/22 LANGPERT sentences obtain scores between 3-5. In terms of mean scores, the 3.12 for NEPERT seems quite different from the overall mean of 2.82.

Additionally, the Table 5 illustrates the inter-annotator disagreement levels for each perturbation category. There is a slight increase from the overall disagreement level ($\delta = 0.61$): for LANGPERT it is 0.73 and for NEPERT 0.72 (for INFLECTPERT close to 0.62). Similarly, the weighted average values of $C\delta_i$ show a significant rise in the disagreement levels of NEPERT (2.16) and LANGPERT (2.5) compared to the overall level (1.85).

If we compare the average unnormalised $C\delta_i$ between with and without the perturbations, we see the average is 2.16 for NEPERT and 1.69 for sentences without NEPERT. Similarly, with INFLECTPERT, the average is 2.2, which is 1.84 without INFLECTPERT, and the averages are 1.85 with and 1.84 without LANGPERT. For NEPERT and INFLECTPERT, the average $C\delta_i$ is considerably different with and without perturbations. Also, we see a reduction in the unanimous agreement $C\delta_i = 0$, which is observed to be 7% on the overall dataset, but for NEPERT it is 3%, for INFLECTPERT is 3.31% and for LANGPERT it's 0% suggesting a higher disagreement in the presence of these perturbations.

To see whether these differences are meaningful, we conduct a multiple linear regression analysis (App. Table 12) between the three perturbation types and closeness scores. Here, three perturbations are the independent variables (one-hot coded),

	Avg. Orig. FRE	Avg. Adv. FRE
Overall	50.91	46.76
NEPERT	49.00	44.00
INFLECTPERT	48.89	44.01
LANGPERT	54.72	48.61

Table 6: Average FRE scores for TRUSTPILOT real and TEXTFOOLER sentences, further separated into perturbation types.

	freq./total (%)	Avg. diff. (↓)
Overall	1008/1800 (56%)	9.49
NEPERT	59/100 (59.00%)	10.25
INFLECTPERT	334/598 (55.85%)	11.29
LANGPERT	15/22 (68.18%)	11.32

Table 7: Frequency of TRUSTPILOT TEXTFOOLER sentences that have lower readability scores than the originals and the average FRE score reduced by the attack.

and their closeness score is the dependent variable. The overall regression is statistically significant, with $p = 0.025 < 0.05$. For the individual perturbation types, only NEPERT demonstrates a significant relationship with the closeness scores (p-value of 0.003). Overall, as expected (given that there are potentially many different types of perturbation), these three types explain only a small proportion of the variation in closeness scores (adjusted R-squared 0.004).

Additionally, a similar analysis conducted on 300 sentences with multiple judgements against the $C\delta_i$ levels (dependent variable), to observe whether these perturbations have any significant impact on the inter-annotator disagreement levels, is presented in App. Table 13. This regression is similarly significant, with a p-value of 0.001, and only NEPERT had a significant coefficient.

Connection with readability. Another method of assessing the effects of these three types of perturbation is to look at changes in an automatic metric. Here, we use the Flesch Reading Ease (FRE) (Flesch, 1979) score, used in text quality assessment to automatically classify text by the level appropriate for learning readers in natural language generation (NLG) (Deutsch et al., 2020; Pitler and Nenkova, 2008) and text summarisation tasks (Luo et al., 2022; Ribeiro et al., 2023).

FRE score (Flesch, 1979) is a readability test designed to indicate how easy or difficult a text is to understand based on the average number of words per sentence and the average number of syllables per word. We use the original formula as derived by Rudolf Flesch (1979):

$$\text{FRE score} = 206.85 - 1.015 \times \frac{\text{total words}}{\text{total sentences}} - 84.6 \times \frac{\text{total syllables}}{\text{total words}} \quad (2)$$

Details on the FRE are in App B. Recalling from adversarial text literature, Garg and Ramakrishnan (2020) argued that in many cases, adversarial algorithms use low-frequency words to modify the original words. A few SOTA adversarial text detection algorithms also rely on the word frequencies in the sentences (Zhou et al., 2019; Mozes et al., 2021). The presence of such infrequent words in sentences might impact the readability level of the sentences as well.

From our investigation of the TRUSTPILOT data, we find in many cases TEXTFOOLER replaces words that decrease the FRE score, making the sentence harder to read, in turn increasing the difference with the real text. Two such example sentences are presented in App. Table 16. Building on this observation, we analyse how much the readability level goes down in the adversarial sentences compared to the original ones. Table 6 illustrates the result. As suspected, the overall average readability level is 50.91 for original sentences and 46.76 for adversarial sentences, with a difference of 4.15. The sentences of the three perturbation groups show a similar trend having lower average FRE for the perturbed sentences than the original ones, for NEPERT sentences the average FRE score of the original sentences is 49 and the adversarial sentences is 44 (difference 5.00), for INFLECTPERT the original average is 48.89 and adversarial average is 44.01 (difference 4.88) and for LANGPERT the average FREs are 54.72 and 48.61 for the original and adversarial sentences, respectively (difference 6.11).

Also as illustrated in Table 7, taking the difference between the readability score of each original-adversarial sentence pair, we observe that for 1008 adversarial sentences out of 1800 sentences (56%) the FRE score is less than the original ones, which is 59 out of 100 sentences (59%) for NEPERT, 334 out of 598 sentences (55.85%) for INFLECTPERT and 15 of 22 sentences (68.18%) for the LANGPERT sentences. For these sentences, where the adversarial readability level is lower than the originals, the average difference in the FRE score is 9.49. If we consider only the sentences having NEPERT, INFLECTPERT and LANGPERT, this difference in

the readability level grows to 10.25, 11.29 and 11.32, respectively, indicating the inclusion of unnatural words.

Considering the effects of selected perturbations on both adversarial closeness scores and readability scores, all the perturbations potentially have notable impacts on the original sentences, so we consider them all for constraints in the adversarial generation process, below.

5 Generating Perturbation-Constrained Adversarial Texts

In this stage, we propose a method TOKENCONSTR to generate better adversarial sentences, by sketching three constraints on the perturbations described in Sec. 4. We add these constraints in addition to the TEXTFOOLER constraints. We then evaluate the adversarial sentences produced by TOKENCONSTR against baseline TEXTFOOLER ones on a new set of human adversarial closeness judgements.

5.1 Experimental Setup

We use the TextAttack (Morris et al., 2020b) framework[7] and TEXTFOOLER as the baseline attack, and add the following additional constraints:

NECONSTR. We restrict the algorithm to reject the candidate sentences that alter the named entities from the original sentence. To detect named entity recognition (NER) changes automatically, we use the SpaCy tagger[8] and restrict ourselves to four types: 'PERSON', 'GPE', 'ORG' and 'LOC'.

INFLECTCONSTR. This constraint is added in two steps. First, using Morpheus (Tan et al., 2020), we include some additional candidate sentences altering the inflection of the original words. Then, we place an additional constraint that rejects a candidate if the perturbed word's POS form does not align with the original word's inflectional POS forms. We use LEMMINFLECT[9] to generate word inflections. App. Table 11 shows an example.

LANGCONSTR. We restrict the modifications of the words to non-English words. Using the LangDetect module[10] part of Google's Language-Detection library, we detect the possible languages for each of the perturbed word. If for a transformed sentence, English is not detected as a language for any of the perturbed words, the constraint doesn't accept the sentence.

Additionally, we apply autocorrect Python spell-checker[11] and convert the sentences to sentence case by using Pytorch-TrueCase library[12] beforehand to maximise the NE recognition.

We only consider TEXTFOOLER sentences from TRUSTPILOT for which human judgement was very poor (adversarial closeness scores of 4 and 5), giving us 449 sentences. We then apply the NECONSTR, INFLECTCONSTR and LANGCONSTR, together referred to as TOKENCONSTR, on them.

TEXTFOOLER Scores	TOKENCONSTR Scores			
	1	2	3	4
4	103	72	16	0
5	11	17	23	5

Table 8: Closeness score frequency of TEXTFOOLER sentences constrained by TOKENCONSTR

Closeness ratings. We again collect human annotations on the TOKENCONSTR sentences' closeness to the originals. Similar to Section 3, we present both the original and the TOKENCONSTR sentences in front of the human judges and ask them to evaluate how close the modified sentence is to the real one on the same 1-5 scale. We compare the ratings of the TOKENCONSTR sentences with the earlier collected ratings of their TEXTFOOLER versions.

A few examples where TOKENCONSTR sentences obtain better closeness scores in human evaluation are in App. Table 17. The first four examples contain named entities that were perturbed by TEXTFOOLER. After adding the constraints, the words "mike"(NE tag: PERSON), "alamo"(NE tag: PERSON), "china"(NE tag: LOC) and "gak" (NE tag: GPE) are constrained by the NECONSTR in those sentences.

5.2 Results and discussion

By implementing TOKENCONSTR we achieve a reduction in the adversarial closeness score for 191 out of 364 sentences previously rated as 4, and 56 out of 85 sentences previously rated as 5, enhancing the quality of 247 sentences in total. A detailed breakdown is given in Table 8, with the highest

[7] https://textattack.readthedocs.io/en/master/
[8] https://tinyurl.com/yc3t6amc
[9] https://lemminflect.readthedocs.io/en/latest/
[10] https://github.com/Mimino666/langdetect
[11] https://github.com/filyp/autocorrect
[12] https://github.com/mayhewsw/pytorch-truecaser

	TEXTFOOLER		TOKENCONSTR	
Score	Freq.	%	Freq.	%
1	176	9.78	196	10.89
2	504	28.00	596	33.11
3	671	37.28	795	44.17
4	364	20.22	182	10.11
5	85	4.72	31	1.72
Avg.	2.82		2.54	

Table 9: Adversarial closeness score distribution of the TEXTFOOLER and TOKENCONSTR sentences.

	TEXTFOOLER		TOKENCONSTR	
Overall δ	0.62		0.58	
	Freq.	%	Freq.	%
$C\delta_i$=0	21	7.00	24	8.00
$C\delta_i$=1	97	32.33	106	35.33
$C\delta_i$=2	104	34.67	109	36.33
$C\delta_i$=3	63	21.00	47	15.67
$C\delta_i$=4	15	5.00	14	4.67
Wgt. avg.	1.84		0.87	

Table 10: Inter-annotator disagreement of TRUSTPILOT TEXTFOOLER and TOKENCONSTR

number of closeness score-4 sentences (103 sentences) re-annotated to a score of 1, and score-5 sentences (23 sentences) to a score of 3, due to the TOKENCONSTR application.

We use TOKENCONSTR adversarial sentences where these improve over TEXTFOOLER ones, and TEXTFOOLER ones otherwise. As shown in Table 9, this enhances the overall adversarial closeness distribution. The average closeness score decreased to 2.54 from 2.82, and the percentages of sentences with scores of 1, 2, and 3 increased significantly to 10.86%, 33.04%, and 44.24%, respectively, suggesting a closer alignment to the original texts.

In terms of inter-annotator disagreement, presented in the Table 10, the average level of disagreement δ decreased to 0.58 from 0.62 for TOKENCONSTR, with increases in the $C\delta_i$ levels of 0 and 1 from 7% to 8% and from 32.33% to 35.33% respectively. However, a substantial degree of disagreement persists, with the highest level of disagreement $C\delta = 4$ at 4.67%, slightly lower than the 5% observed for TEXTFOOLER.

We observe that these improvements in human-judged closeness — where in 247/449 cases TOKENCONSTR sentences show better closeness to the original ones than TEXTFOOLER— there is no clear relationship to automated readability scoring. Among these 247 sentences, 135 TEXTFOOLER instances have lower readability scores than originals, with an average FRE score reduction of 9.18. TOKENCONSTR also obtains lower readability in 116 cases compared to the originals, with an average readability reduction of 8.57.

6 Conclusion and future work

In this work, we deepened the exploration of human perceptions of adversarial sentences, extending beyond isolated suspiciousness to look at how humans perceive closeness between original sentences and their adversarial alternatives. We then constructed two datasets of these human judgements, giving the judges real and adversarial text pairs and highlighting the perturbed words. We showed that, as with the human suspicion dataset of Tonni et al. (2025), the datasets are of reasonable reliability and capture some interesting insights into human perception of adversarial examples.

Then we identified three types of perturbations — altered named entities (NEPERT), wrong inflectional parts of speech (POS) of the original words (INFLECTPERT), and replacement of words to those from another language (LANGPERT) — and examined their connection to adversarial closeness, and also to an automatic metric measuring reading difficulty. We found that human perception of sentence authenticity is significantly related to at least the NEPERT perturbations. We then added constraints based on these perturbations to generate improved adversarial sentences, with the effectiveness of these adaptations confirmed through further human evaluations.

Future work could explore how humans perceive LLM-generated adversarial sentences, especially those designed to preserve author anonymity through stylometric considerations (Kandula et al., 2024; Fisher et al., 2024; Staab et al., 2025). The naturalness and imperceptibility of the adversarial sentences produced by LLMs have already been considered a crucial factor by Xu and Wang (2024). Additionally, Faustini et al. (2025) found that LLMs often produce texts that vary significantly from the original ones. Thus, evaluating the adversarial closeness of such generated sentences and applying the TOKENCONSTR in LLM prompts in sentence generation would also be valuable. Also, further work on the automated prediction of human perception of adversarial closeness and investigation of the connection between readability and human perception may reveal more insights into adversarial text quality.

7 Limitation

This study primarily examines perturbations introduced by TEXTFOOLER, a single adversarial attack. Extending this analysis to other adversarial attacks and domains could identify additional perturbations that significantly influence human perception of adversarial sentences. Furthermore, exploring ways to predict human perception of adversarial closeness by the language models can help to build a useful automated adversarial closeness scorer besides human evaluation.

8 Acknowledgement

This project was undertaken with the assistance of resources and services from the National Computational Infrastructure (NCI), supported by the Australian Government. This project was also supported by the International Macquarie University Research Excellence Scholarship. The human evaluation section of this study has received ethics approval from Macquarie University (Human Ethics Comm. Approval Code: 5201800393).

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, Brussels, Belgium. Association for Computational Linguistics.

Nicholas Carlini and David A. Wagner. 2017. Towards evaluating the robustness of neural networks. In *2017 IEEE Symposium on Security and Privacy, SP 2017, San Jose, CA, USA, May 22-26, 2017*, pages 39–57. IEEE Computer Society.

Yize Cheng, Vinu Sankar Sadasivan, Mehrdad Saberi, Shoumik Saha, and Soheil Feizi. 2025. Adversarial paraphrasing: A universal attack for humanizing ai-generated text. *CoRR*, abs/2506.07001.

Ryan Andrew Chi, Nathan Kim, Patrick Liu, Zander Lack, and Ethan A Chi. 2022. GLARE: Generative left-to-right AdversaRial examples. In *Proceedings of the 3rd Workshop on Evaluation and Comparison of NLP Systems*, pages 44–50, Online. Association for Computational Linguistics.

Tovly Deutsch, Masoud Jasbi, and Stuart Shieber. 2020. Linguistic features for readability assessment. In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 1–17, Seattle, WA, USA → Online. Association for Computational Linguistics.

Salijona Dyrmishi, Salah Ghamizi, and Maxime Cordy. 2023. How do humans perceive adversarial text? a reality check on the validity and naturalness of word-based adversarial attacks. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8822–8836, Toronto, Canada. Association for Computational Linguistics.

Javid Ebrahimi, Anyi Rao, Daniel Lowd, and Dejing Dou. 2018. HotFlip: White-box adversarial examples for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 31–36, Melbourne, Australia. Association for Computational Linguistics.

Steffen Eger and Yannik Benz. 2020. 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*, pages 786–803, Suzhou, China. Association for Computational Linguistics.

Pedro Faustini, Shakila Mahjabin Tonni, Annabelle McIver, Qiongkai Xu, and Mark Dras. 2025. Idt: Dual-task adversarial rewriting for attribute anonymization. *Computational Linguistics*, pages 1–39.

Jillian Fisher, Skyler Hallinan, Ximing Lu, Mitchell L Gordon, Zaid Harchaoui, and Yejin Choi. 2024. StyleRemix: Interpretable authorship obfuscation via distillation and perturbation of style elements. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4172–4206, Miami, Florida, USA. Association for Computational Linguistics.

Rudolf Flesch. 1979. *How to write plain English: A book for lawyers and consumers*, volume 76026225. Harper & Row New York.

Ji Gao, Jack Lanchantin, Mary Lou Soffa, and Yanjun Qi. 2018. Black-box generation of adversarial text sequences to evade deep learning classifiers. In *2018 IEEE Security and Privacy Workshops, SP Workshops 2018, San Francisco, CA, USA, May 24, 2018*, pages 50–56. IEEE Computer Society.

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, Online. Association for Computational Linguistics.

Dirk Hovy, Anders Johannsen, and Anders Søgaard. 2015. User review sites as a resource for large-scale sociolinguistic studies. In *Proceedings of the 24th International Conference on World Wide Web, WWW 2015, Florence, Italy, May 18-22, 2015*, pages 452–461. ACM.

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 *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 8018–8025. AAAI Press.

Hemanth Kandula, Damianos Karakos, Haoling Qiu, and Brian Ulicny. 2024. Improving authorship privacy: Adaptive obfuscation with the dynamic selection of techniques. In *Proceedings of the Fifth Workshop on Privacy in Natural Language Processing*, pages 137–142, Bangkok, Thailand. Association for Computational Linguistics.

Peter Kincaid, Robert P. Fishburne, Richard L. Rogers, and Brad S. Chissom. 1975. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel.

Dianqi Li, Yizhe Zhang, Hao Peng, Liqun Chen, Chris Brockett, Ming-Ting Sun, and Bill Dolan. 2021. 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*, pages 5053–5069, Online. Association for Computational Linguistics.

Guoyi Li, Bingkang Shi, Zongzhen Liu, Dehan Kong, Yulei Wu, Xiaodan Zhang, Longtao Huang, and Honglei Lyu. 2023. Adversarial text generation by search and learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 15722–15738, Singapore. Association for Computational Linguistics.

Jinfeng Li, Shouling Ji, Tianyu Du, Bo Li, and Ting Wang. 2019. Textbugger: Generating adversarial text against real-world applications. In *26th Annual Network and Distributed System Security Symposium, NDSS 2019, San Diego, California, USA, February 24-27, 2019*. The Internet Society.

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, Online. Association for Computational Linguistics.

Han Liu, Zhi Xu, Xiaotong Zhang, Feng Zhang, Fenglong Ma, Hongyang Chen, Hong Yu, and Xianchao Zhang. 2023. Hqa-attack: Toward high quality black-box hard-label adversarial attack on text. In *Advances in Neural Information Processing Systems*, volume 36, pages 51347–51358. Curran Associates, Inc.

Zheheng Luo, Qianqian Xie, and Sophia Ananiadou. 2022. Readability controllable biomedical document summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4667–4680, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Xingjun Ma, Bo Li, Yisen Wang, Sarah M. Erfani, Sudanthi N. R. Wijewickrema, Grant Schoenebeck, Dawn Song, Michael E. Houle, and James Bailey. 2018. Characterizing adversarial subspaces using local intrinsic dimensionality. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

John Morris, Eli Lifland, Jack Lanchantin, Yangfeng Ji, and Yanjun Qi. 2020a. Reevaluating adversarial examples in natural language. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3829–3839, Online. Association for Computational Linguistics.

John Morris, Eli Lifland, Jin Yong Yoo, Jake Grigsby, Di Jin, and Yanjun Qi. 2020b. 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*, pages 119–126, Online. Association for Computational Linguistics.

Maximilian Mozes, Pontus Stenetorp, Bennett Kleinberg, and Lewis Griffin. 2021. Frequency-guided word substitutions for detecting textual adversarial examples. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 171–186, Online. Association for Computational Linguistics.

Bo Pang and Lillian Lee. 2005. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 115–124, Ann Arbor, Michigan. Association for Computational Linguistics.

Emily Pitler and Ani Nenkova. 2008. Revisiting readability: A unified framework for predicting text quality. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 186–195, Honolulu, Hawaii. Association for Computational Linguistics.

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, Florence, Italy. Association for Computational Linguistics.

Shuhuai Ren, Yihe Deng, Kun He, and Wanxiang Che. 2019. Generating natural language adversarial examples through probability weighted word saliency. In

Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 1085–1097, Florence, Italy. Association for Computational Linguistics.

Leonardo F. R. Ribeiro, Mohit Bansal, and Markus Dreyer. 2023. Generating summaries with controllable readability levels. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11669–11687, Singapore. Association for Computational Linguistics.

Yundi Shi, Piji Li, Changchun Yin, Zhaoyang Han, Lu Zhou, and Zhe Liu. 2022. Promptattack: Prompt-based attack for language models via gradient search. In *Natural Language Processing and Chinese Computing - 11th CCF International Conference, NLPCC 2022, Guilin, China, September 24-25, 2022, Proceedings, Part I*, volume 13551 of *Lecture Notes in Computer Science*, pages 682–693. Springer.

Robin Staab, Mark Vero, Mislav Balunovic, and Martin Vechev. 2025. Language models are advanced anonymizers. In *The Thirteenth International Conference on Learning Representations*.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, and Rob Fergus. 2014. Intriguing properties of neural networks. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*.

Samson Tan, Shafiq Joty, Min-Yen Kan, and Richard Socher. 2020. It's morphin' time! Combating linguistic discrimination with inflectional perturbations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2920–2935, Online. Association for Computational Linguistics.

Shakila Mahjabin Tonni, Pedro Faustini, and Mark Dras. 2025. Graded suspiciousness of adversarial texts to humans. *Computational Linguistics*, 51(3):705–738.

Carl Vogel, Maria Koutsombogera, and Rachel Costello. 2020. Analyzing likert scale inter-annotator disagreement. In Anna Esposito, Marcos Faúndez-Zanuy, Francesco Carlo Morabito, and Eros Pasero, editors, *Neural Approaches to Dynamics of Signal Exchanges*, volume 151 of *Smart Innovation, Systems and Technologies*, pages 383–393. Springer.

Tianlu Wang, Xuezhi Wang, Yao Qin, Ben Packer, Kang Li, Jilin Chen, Alex Beutel, and Ed Chi. 2020. CAT-gen: Improving robustness in NLP models via controlled adversarial text generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5141–5146, Online. Association for Computational Linguistics.

Zimu Wang, Wei Wang, Qi Chen, Qiufeng Wang, and Anh Nguyen. 2024. Generating valid and natural adversarial examples with large language models. In *27th International Conference on Computer Supported Cooperative Work in Design, CSCWD 2024, Tianjin, China, May 8-10, 2024*, pages 1716–1721. IEEE.

Yue Xu and Wenjie Wang. 2024. LinkPrompt: Natural and universal adversarial attacks on prompt-based language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6473–6486, Mexico City, Mexico. Association for Computational Linguistics.

Eray Yildiz and A. Cüneyd Tantuğ. 2019. Morpheus: A neural network for jointly learning contextual lemmatization and morphological tagging. In *Proceedings of the 16th Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 25–34, Florence, Italy. Association for Computational Linguistics.

Jin Yong Yoo and Yanjun Qi. 2021. Towards improving adversarial training of NLP models. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 945–956, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yichao Zhou, Jyun-Yu Jiang, Kai-Wei Chang, and Wei Wang. 2019. Learning to discriminate perturbations for blocking adversarial attacks in text classification. 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)*, pages 4904–4913, Hong Kong, China. Association for Computational Linguistics.

Ying Zhou, Ben He, and Le Sun. 2024. Humanizing machine-generated content: Evading AI-text detection through adversarial attack. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 8427–8437, Torino, Italia. ELRA and ICCL.

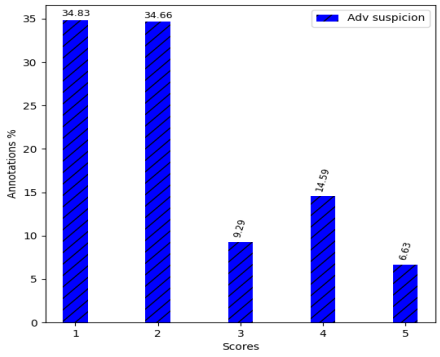
Andy Zou, Zifan Wang, J. Zico Kolter, and Matt Fredrikson. 2023. Universal and transferable adversarial attacks on aligned language models. *CoRR*, abs/2307.15043.

Appendix

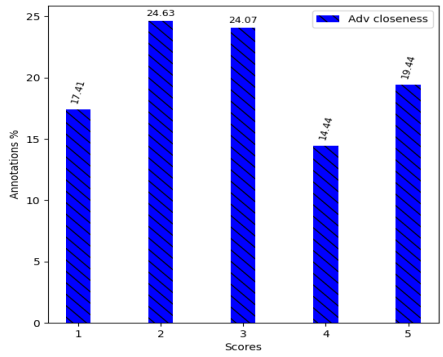
A Adversarial closeness vs. suspiciousness.

Comparing human evaluations for MOVIEREVIEW TEXTFOOLER suspiciousness and adversarial closeness scores (illustrated in the Figure 1). we get a substantial correlation with Pearson's correlation coefficient of 0.58, suggesting that they are fairly closely aligned. However, in case of suspiciousness, the human judges generally avoided choosing a score of 3 to indicate "uncertain", while for adversarial closeness judges choose a score of 3, as a rank, quite freely.

So, removing the scores of 3, implying a definite opinion (scores of $1-2$ and $4-5$), we see the scores are skewed towards the non-suspicion with 69.49% of the sentences to be less suspicious (scored 1 and 2) and 21.22% to be more suspicious (scored 4 and 5), which varies drastically for closeness. In adversarial closeness, only 41.96% sentences are deemed to resemble the originals (scores of 1 and 2) and 33.88% to be very different (scores of 4 and 5).



(a)



(b)

Figure 1: Human judgements (%) of a) human suspiciousness and b) adversarial closeness to the original sentence of MOVIEREVIEW

B Flesch Reading Ease (FRE)

It's calculated using average sentence length and the average number of syllables per word, with polysyllabic words having a larger impact than on grade-level scores. The score ranges from 0 to 100, with higher scores indicating easier readability and lower scores indicating more complex material.

Flesch Reading Ease (FRE) scores, developed by Flesch (1979) (originally proposed in 1940) while working with the Associated Press as a method for improving the readability of newspapers.

The score indicates how easy or difficult a piece of text is to understand, based on a formula that considers average sentence length and the average number of syllables per word. Thus, polysyllabic words affect this score significantly more than they do the grade-level score. This score is interpreted on a scale where higher scores indicate material that is easier to read, and lower scores indicate more complex material.

The score typically ranges from 0 to 100, with higher scores indicating easier readability. The score can be interpreted as follows:

- 90-100: Very Easy to read. Easily understood by an average 11-year-old student.
- 80-89: Easy to read. Conversational English for consumers.
- 70-79: Fairly easy to read.
- 60-69: Plain English. Easily understood by 13- to 15-year-old students.
- 50-59: Fairly difficult to read.
- 30-49: Difficult to read.
- 0-29: Very difficult to read. Best understood by university graduates.

Later, Kincaid et al. (1975) derives the following adjusted version of the FRE score by conducting multiple regression analysis for the U.S. Navy:

$$\text{Flesh-Kincaid readability score} = 0.39 \times \frac{\text{total words}}{\text{total sentences}} + 11.8 \times \frac{\text{total syllables}}{\text{total words}} - 15.59 \quad (3)$$

However, in our experiments, we found that the adversarial perturbations have a greater impact on the FRE scores than on the Flesch-Kincaid readability scores. Thus, we use FRE as an automated metric to assess the effect of the perturbations.

C TOKENCONSTR Sentence Generation

An example original sentence and its TOKENCONSTR sentence generation using the *TextAttack* framework is illustrated in Table 11.

Original	had tyre fitted at national tyres in skelmersdale was in and out within mins. excellent service will recommend and certainly use again
Transformation Tan et al. (2020)	had tyre ['fits', 'fit', 'fitting'] at national ['tyre'] in skelmersdale was in and out within mins. excellent ['services', 'serviced','servicing'] will recommend and certainly use again
Constraint	Orig. token: service Orig. inflections: ['services', 'service', 'service', 'serviced', 'servicing' 'services', 'service', 'service'] Orig. inflection POS: ['NNS', 'NN', 'VBD', 'VBG', 'VBZ', 'VB', 'VBP'] Pert. token: servicing Pert. inflection POS:'VBG'
	Orig. token: excellent Orig. inflections:['excellent'] Orig. inflection POS: ['JJ'] Pert. token: marvelous Pert. inflection POS: 'JJ'
INFLECTCONSTR	had wheeled adjusting at national tyres in skelmersdale was in and out within mins. marvelous servicing will recommend and certainly use again

Table 11: TRUSTPILOT example applying two-step INFLECTCONSTR- as a **transformation** following Morpheus (Tan et al., 2020) and as a **constraint**

D MTurk UI and HIT Setup

The MTurk user interface with an example sentence-pair is shown in Figure 2. We present each pair of sentences as a single HIT and ask to score how close the "computer-altered" sentence is to the "human-written" one. Mturkers were paid A$0.13/HIT, which is the standard rate of A$0.20/HIT. To control the quality of the collected annotations, we only choose English-speaking workers with an MTurk Master's level qualification and HIT approval rate above 95%. The Amazon Mechanical Turk user interface is illustrated in Figure 2.

E Multiple Linear Regression Analysis

The results of multiple linear regression analysis between the three perturbations and their adversarial closeness score are reported in Table 12 and the results of the analysis between the perturbations and the inter-annotator disagreement in Table 13.

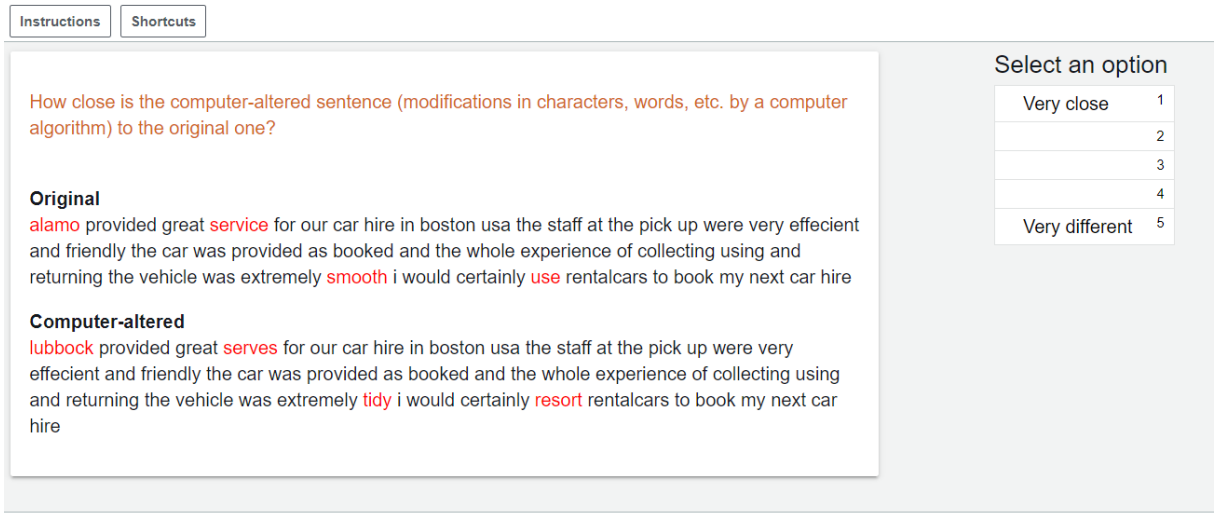


Figure 2: MTurk UI with a TRUSTPILOT sentence for the adversarial closeness judgement

Regression Statistics								
Multiple R 0.072		R Square 0.005		Adjusted R Square 0.004		Std. Error 1.013		
ANOVA								
	df	SS	MS	F	Sig. F			
Regression	3	9.647	3.216	3.131	0.025			
Residual	1796	1844.751	1.027					
Total	1799	1854.398						
	Coeff.	Std. Error	t Stat	P-val.	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	2.800	0.030	94.308	0	2.742	2.858	2.742	2.858
NEPERT	0.316	0.105	3.020	0.003	0.111	0.521	0.111	0.521
INFLECTPERT	0.013	0.051	0.253	0.800	-0.086	0.113	-0.087	0.113
LANGPERT	-0.074	0.218	-0.340	0.734	-0.501	0.353	-0.501	0.353

Table 12: Multiple linear regression output for the three perturbations against the adversarial closeness

F Adversarial Examples

MOVIEREVIEW sentences attacked by TEXTFOOLER and along with obtained human judgements on the adversarial closeness scores and suspiciousness score according to Tonni et al. (2025) are depicted in Table 14. Similarly, Table 4 illustrates examples on TRUSTPILOT. Table 17 contains example sentences showing both TEXTFOOLER and TOKENCONSTR versions and their closeness scores. We also give examples of how readability differs between real and TEXTFOOLER sentences on TRUSTPILOT in Table 16.

Regression Statistics								
Multiple R 0.225		R Square 0.051		Adjusted R Square 0.041		Std. Error 0.976		
ANOVA								
	df	SS	MS	F	Sig. F			
Regression	3.000	15.067	5.022	5.274	0.001			
Residual	296.000	281.879	0.952					
Total	299.000	296.947						
	Coeff.	Std. Error	t Stat	P-val.	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	1.700	0.081	20.869	0.000	1.539	1.860	1.539	1.860
NEPERT	0.469	0.120	3.891	0.001	0.232	0.706	0.232	0.706
INFLECTPERT	-0.032	0.116	-0.281	0.779	-0.260	0.195	-0.260	0.195
LANGPERT	0.225	0.442	0.510	0.610	-0.645	1.096	-0.645	1.096

Table 13: Multiple linear regression output for the three perturbations against the $C\delta_i$

Text	Pert. Type	Suspiciousness (Tonni et al., 2025)	Closeness
orig. — a markedly inactive film city is conversational bordering on confessional. **adv.** — a markedly idling film city is conversational bordering on confessional.	INFLECTPERT	2 4	1
orig. — what the four feathers lacks is genuine sweep or feeling or even a character worth caring about. **adv.** —what the four feathers rarity is genuine sweep or feel or even a character worth caring about.	INFLECTPERT	1 3	1
orig. — frida is certainly no disaster but neither is it the kahlo movie frida fans have been looking for. **adv.** — frida is visibly no disaster but neither is it the kahlo stills freda fans have been looking for.	NEPERT	1 5	5
orig. — the movie would seem less of a trifle if ms. sugarman followed through on her defiance of the saccharine. **adv.** — the movie would seems less of a trifle if tatjana. sugarman followed through on her defiance of the saccharine.	NEPERT	2 4	4
orig. — crystal and de niro manage to squeeze out some good laughs but not enough to make this silly con job sing. **adv.** — crystalline and las pesci manage to squeeze out some nice amused but not enough to make this nutty con mission sing.	NEPERT / LANGPERT	1 5	5
orig. — boasts eye-catching art direction but has a forcefully quirky tone that quickly wears out its limited welcome. **adv.** — boasts eye-catching art direction but has a forcefully quirky tone that quickly porte out its limited welcome.	LANGPERT	2 5	3

Table 14: MOVIEREVIEW sentences with adversarial perturbations by TEXTFOOLER- changes in the morphological inflection (INFLECTPERT), changes in named entities (NEPERT) and spurious perturbations from or to non-English words (LANGPERT) along with obtained human judgements on the adversarial closeness scores on them. We also report the suspiciousness scores from Tonni et al. (2025)

Text	Pert. Type	Closeness
orig. — live chat feature enabled me to make an informed purchase after speaking with a technical salesperson cheers mike **adv.** — live chat mannerisms empowering me to make an informed purchase after speaking with a technician salesperson cheers michaela	INFLECTPERT	3
orig. — every time I've ordered something from Playcom something has gone wrong and that covers the last five years stick to Amazon or eBay **adv.** — every time i've ordered something from playcom something has gone amiss and that blanket the iast five aged stick to amazon or ebay	INFLECTPERT	4
orig. — i bought gb of compatible ram from crucialcom based on their system scanner software **adv.** — i bought megs of obedient ramallah from crucialcom reasoned on their programmes scanning sw	NEPERT	4
orig. — having spent a lot of time and money importing gbics and dac cables from the us and china i decided to leave the hassle **adv.** — having spent a lot of time and money importing gbics and dac yarn from the us and porcelain i decided to leave the hassle	NEPERT	5
orig. — a saving of over is anticipated against my next full years bill **adv.** — a saving of over is anticipated against my next holistic aged lois	LANGPERT	4
orig. — got what i wanted at a good price and came straight away pucker **adv.** — got what i wanted at a lovely costa and came straight away pouty	LANGPERT	3

Table 15: TRUSTPILOT TEXTFOOLER generated sentences with INFLECTPERT, NEPERT and LANGPERT perturbations and their closeness scores (lower is better).

Text	FRE Score
orig. — good quality and very cheap capo that does what it should very quick delivery as well	63.70
adv. — good quality and exceedingly affordable capo that does what it would extraordinarily speedy delivery as also *(closeness:3)*	5.53
orig. — arrived on time less than hrs what more can i say will use again	89.90
adv. — took on time cheaper than afternoon what more can i explaining yearning used again *(closeness:5)*	59.68

Table 16: FRE readability scores TRUSTPILOT original and TEXTFOOLER sentences along with their closeness scores.

Orig. #1 live chat feature enabled me to make an informed purchase after speaking with a technical salesperson cheers mike
TEXTFOOLER *(closeness:3) #* live chat mannerisms empowering me to make an informed purchase after speaking with a technician salesperson cheers michaela TOKENCONSTR *(closeness:2) #* live chat feature helps me to make an informed purchase after speaking with a technical marchand cheerfulness mike
Orig. #2 alamo provided great service for our car hire in boston usa the staff at the pick up were very effecient and the whole experience of collecting using and returning the vehicle was extremely smooth
TEXTFOOLER *(closeness:3) #* lubbock provided great serves for our car hire in boston usa the staff at the pick up were very effecient and the whole experience of collecting using and returning the vehicle was extremely tidy TOKENCONSTR *(closeness:2) #* alamo provided wonderful serves for our car hire in boston usa the servants at the pick up were very efficient and the whole experience of collecting using and returning the vehicle was extremely smoothly
Orig. #3 having spent a lot of time and money importing gbics and dac cables from the us and china i decided to leave the hassle
TEXTFOOLER *(closeness:5) #* having spent a lot of time and money importing gbics and dac yarn from the us and porcelain i decided to leave the hassle TOKENCONSTR *(closeness:3) #* having spent a lot of time and money importing gbics and dac telegrams from the us and china u decided to leave the hassle
Orig. #4 gak used before always on the short list for places to buy from delivered on time easy website and checkout
TEXTFOOLER *(closeness:5) #* cuma used before always on the short list for places to buy from delivered on time easy website and lookat TOKENCONSTR *(closeness:3) #* gak used before always on the little list for places to buy from dispatched on time easy website and lookat
Orig. #5 dude this place is great for scrapping any unwanted computor and they give you a quote no strings attached
TEXTFOOLER *(closeness:4) #* matti this placing is great for remove any unwanted computor and they give you a quote no fetters attached TOKENCONSTR *(closeness:3) #* mate this place is great for scrapping any unwanted machine and they give you a quote no strings attached

Table 17: TRUSTPILOT and TOKENCONSTR sentences and obtained human closeness scores on them.