Graded Suspiciousness of Adversarial Texts to Humans

Shakila Mahjabin Tonni[1], Pedro Faustini[2], and Mark Dras[2]

[1]Macquarie University / School of Computing; Data61, CSIRO
shakila.tonni@data61.csiro.au
[2]Macquarie University, School of Computing

Adversarial examples pose a significant challenge to deep neural networks across both image and text domains, with the intent to degrade model performance through carefully altered inputs. Adversarial texts, however, are distinct from adversarial images due to their requirement for semantic similarity and the discrete nature of the textual contents. This study delves into the concept of human suspiciousness, a quality distinct from the traditional focus on imperceptibility found in image-based adversarial examples, where adversarial changes are often desired to be indistinguishable to the human eye even when placed side by side with originals. Although this is generally not possible with text, textual adversarial content must still often remain undetected or non-suspicious to human readers. Even when the text's purpose is to deceive NLP systems or bypass filters, the text is often expected to be natural to read.

In this research, we expand the study of human suspiciousness by analyzing how individuals perceive adversarial texts. We gather and publish a novel dataset of Likert-scale human evaluations on the suspiciousness of adversarial sentences, crafted by four widely used adversarial attack methods and assess their correlation with the human ability to detect machine-generated alterations. Additionally, we develop a regression-based model to predict levels of suspiciousness and establish a baseline for future research in reducing the suspiciousness in adversarial text generation. We also demonstrate how the regressor-generated suspicious scores can be incorporated into adversarial generation methods to produce texts that are less likely to be perceived as computer-generated.

1. Introduction

Adversarial examples, which maliciously aim at degrading the performance of deep neural networks by carefully crafted inputs, have shown to be effective in both image and text domains (Papernot et al. 2016; Carlini and Wagner 2017; Ma et al. 2018; Moosavi-Dezfooli, Fawzi, and Frossard 2016; Pruthi, Dhingra, and Lipton 2019; Alzantot et al. 2018; Li et al. 2020; Garg and Ramakrishnan 2020, for example).

To generate adversarial examples, inputs for which the model's predictions are correct are targeted and modified slightly, causing the model to alter its prediction.

In the case of an image, this can be accomplished by altering only a few pixels (Szegedy et al. 2014; Carlini and Wagner 2017). However, generating adversarial texts

---

Action Editor: Rico Sennrich. Submission received: 7 June 2024; revised version received: 23 November 2024; accepted for publication: 9 January 2025.

https://doi.org/10.1162/coli_a_00555

© 2025 Association for Computational Linguistics
Published under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International (CC BY-NC-ND 4.0) license

differs in a number of ways: Input is discrete, there is often a requirement to be semantically similar to the original, and so forth. Adversarial texts are often generated by modifying the sentences either at the character level by adding, deleting, or altering one or more characters from a sentence (Li et al. 2019; Pruthi, Dhingra, and Lipton 2019; Eger and Benz 2020) or at the word level by altering a few of the words with semantic preserving synonymous words or tokens (Jin et al. 2020; Garg and Ramakrishnan 2020; Alzantot et al. 2018).

In adversarial images, the added noises are mostly very subtle and imperceptible to humans, with this often being an explicit goal or criterion. However, there are ongoing discussions about what an analogous condition is for textual adversarial examples, and how necessary it is. Some earlier studies (Li et al. 2020) identify imperceptibility as an explicit objective, but without providing a precise definition. This goal is straightforward for images, where the original and adversarial versions can be displayed side by side for human judgment on whether any difference exists. However, this approach is not feasible for text as the differences are typically visible and obvious.

Chen et al. (2022) argued that, based on realistic scenarios where adversarial examples might be deployed, imperceptibility is not a necessary criterion, giving as an example an attempt to bypass an offensive content filter on the Web while still conveying the offensive meaning.

However, there are human-in-the-loop scenarios where human impressions of the text are important. Consider reviews posted to a Web site that aim to bypass a filter or fool a classifier but that shouldn't be dismissed by human readers as obviously computer-generated or computer-altered. Alternatively, phishing emails have to evade spam filters but still look plausible to humans: The standard advice in cybersecurity courses is for the recipient to look for errors and other signs of poor-quality text.[1]

These sorts of scenarios led Morris et al. (2020a), in extending to NLP the ontology of attack spaces proposed by Gilmer et al. (2018), to propose a notion of **(non-)suspicion**; this is one of four kinds of constraints (along with semantics, grammaticality, and overlap) that are proposed to be applied to adversarial examples. Non-suspicion differs somewhat from imperceptibility, in that it focuses on an awareness by a human reader that a sentence may have been modified. Dyrmishi, Ghamizi, and Cordy (2023) adopt this as part of their extensive analysis of adversarial example methods in NLP:

> For the many NLP systems that interact with humans, we argue that effective adversarial attacks should produce valid and natural text. Validity refers to the property humans perceive as having the same semantic properties of interest for an adversarial text as for the original text from which it was produced. Naturalness refers to the perception that humans produce an adversarial text. Adversarial texts that are invalid and/or unnatural can still cause failed NLP model decisions, however, their ultimate effect on humans is negligible because they would fail to convey the intended meaning (e.g. hate speech that is not perceived as hateful) or they would be suspected to be computer-generated (e.g., a phishing email using awkward vocabulary and grammar).

However, both Morris et al. (2020a) and Dyrmishi, Ghamizi, and Cordy (2023) only consider non-suspicion in a binary manner.[2] It is clear that there are gradations in the

---

[1] `https://www.cisa.gov/news-events/news/avoiding-social-engineering-and-phishing-attacks`.

[2] Dyrmishi, Ghamizi, and Cordy (2023), in the above quote, use the term "naturalness," by which they are referring to a concept that strictly subsumes suspiciousness. They note that in addition to evaluating the suspiciousness notion from Morris et al. (2020a), they investigate three related criteria under the umbrella of naturalness. In this paper, we consider only suspiciousness.

Table 1
A human-authored movie review with two adversarial variants generated by #1 PRUTHI (Pruthi, Dhingra, and Lipton 2019) and #2 TEXTFOOLER (Jin et al. 2020). Altered words are in blue.

original	graced with the kind of social texture and realism that would be foreign in american teen comedies
adv #1	gaced with the kind of social texture and realism that would be foreign in american teen comedies
adv #2	complied with the kind of social textures and lifelike that would be foreign in usa childish comedies

suspiciousness with which a human might view an adversarial example. Consider the examples in Table 1, with an original movie review and two adversarial variants of it. The first adversarial variant might not trigger much suspicion, a human considering that it could have been caused by a human typo; the second is much more suspicious. Moreover, Dyrmishi, Ghamizi, and Cordy (2023) do compile statistics aggregating their binary human judgments and then use these statistics as evidence for some adversarial methods producing more suspicious texts than others in the aggregate. Having these graded suspiciousness annotations for *individual* items permits a much wider range of possible uses than binary ones: building scorers that can assess relative likelihoods of being computer-altered, incorporating suspiciousness as a constraint in generating adversarial examples, and so on.[3]

In this work, we present an investigation of human judgments of the suspiciousness of adversarial texts that goes beyond the binary of the above work to reflect gradations of suspiciousness, such as might reflect differences in individual texts like those in Table 1. First, we collect a dataset of Likert-scale suspiciousness judgments on adversarial sentences generated by four widely used methods (Pruthi, Dhingra, and Lipton 2019; Alzantot et al. 2018; Jin et al. 2020; Garg and Ramakrishnan 2020), and analyze some characteristics of this dataset in light of observations of human accuracy in detecting machine-generated text (Ippolito et al. 2020; Clark et al. 2021). Second, as a baseline for future work on this task, we build a regression-based scorer from these judgments and evaluate its reliability on held-out texts as a method for predicting the suspiciousness of adversarial texts. Third, as an illustration of an application of such a scorer, we carry out a small-scale experiment to assess the usefulness of such a scorer in improving suspiciousness in generating adversarial texts.

The contributions of this article are then as follows:

- We make available our novel dataset of Likert-scale human judgments on the suspiciousness of human and adversarial texts, along with an analysis of this data and a discussion of how this fits with prior work.[4]

---

[3] We observe that there are similar approaches to related tasks like deepfake video and audio detection (Sundar, Molina, and Cho 2021; Mai et al. 2023; Groh et al. 2024), which have moved beyond binary judgments to graded Likert-scale judgments of individual items for various purposes: understanding levels of confidence in perception or helping direct human judges to "appropriately allocate attention to the content or perceptual cues" (Groh et al. 2024).

[4] Annotated data and code: `https://github.com/SJabin/HumanAdvSuspiciousness`.

- We define and compare a range of methods for producing regression-based scorers for predicting suspicion judgments, and determine the most effective.
- We demonstrate that such scorers can be integrated into adversarial text generation methods to generate less suspicious-looking texts.

2. Related Works

2.1 Crafting Adversarial Sentences

Adversarial texts are generated by modifying part of an input sentence so that the output becomes incorrect (Goodfellow, Shlens, and Szegedy 2015). This is an area of active interest in NLP, with several surveys providing overviews of techniques (Zhang, Sheng, et al. 2020; Han et al. 2022, for example); we briefly discuss relevant work below.

The most common type of model attack is against classifiers, aiming to change the predicted class (Zhang, Sheng, et al. 2020), and the most common of these kinds of adversarial texts introduce perturbations at the character level or the word level. These are the ones we focus on in this article, although we briefly discuss others for completeness. Many of these types of adversarial attacks have been facilitated by the *TextAttack* platform (Morris et al. 2020b), which frames them as specific instances of a more general combinatorial optimization task.

Character-level Attacks. Character-level attacks alter the characters within a sentence (Pruthi, Dhingra, and Lipton 2019; Li et al. 2019; Gao et al. 2018; Ebrahimi et al. 2018, for example) by adding, removing, or exchanging characters. These techniques often result in adversarial examples that lack grammatical correctness and may easily be identified by humans, if too unnatural.

Word-level Attacks. Word-level attacks, on the other hand, focus on replacing words or tokens of a sentence with a synonymous word (Garg and Ramakrishnan 2020; Jin et al. 2020; Ren et al. 2019; Alzantot et al. 2018; Ribeiro et al. 2020, for example). Among the word-level attacks, Alzantot et al. (2018) and Jia et al. (2019) ran a generic algorithm to generate adversarial text. PWWS (Ren et al. 2019), Particle Swarm Optimization (Zang et al. 2020), and TextFooler (Jin et al. 2020) transform texts with synonyms from lexical databases and word embeddings.

More recent word-level attacks have focused more on perturbation with improved contexts and take benefit from the transformer-based masked language models (MLMs). A2T (Yoo and Qi 2021) uses MLM word replacement to train transformer-based models (BERT, RoBERTa). BERT-Attack (Li et al. 2020) and BAE (Garg and Ramakrishnan 2020) used both word insertion and replacement with the MLM. CLARE (Li et al. 2021) allows replacing, inserting, and merging of tokens to produce variable-length adversarial texts. GLARE (Chi et al. 2022) applies generative models (GPT-2) for text infilling. CAT-Gen (Wang et al. 2020) pre-trains an attribute classifier to alter a token from one attribute group by a token from another attribute group. TextCheater (Peng et al. 2023) proposes a metaheuristic query-efficient search algorithm and LimeAttack (Zhu et al. 2024) uses a local explainable method to generate adversarial texts in the hard-label context. SSCAE (Asl et al. 2024) imposes thresholds on Universal Sentence Encoder-based and

GPT-2-based semantic and syntactic correctness to generate imperceptible adversarial examples. There are a few successful attempts to attack large language models (LLMs) as well (Zhang et al. 2024; Wang et al. 2023a,b).

Phrase-level Attacks. A (much smaller) third category consists of phrase-level or paraphrastic attacks. Iyyer et al. (2018) introduced syntactically controlled paraphrase networks that create paraphrased adversarial sentences using a distinct constituency-based syntax format. Qi et al. (2021) developed a style transfer–based adversarial attack by repeatedly transforming original inputs into various text styles to generate adversarial examples; for example, the register of the text could be changed from social media posts to biblical. The Phrase-Level Textual Adversarial Attack (Lei et al. 2022) generates adversarial samples by modifying vulnerable phrases identified as attack targets by a syntactic parser, employing a pre-trained blank-infilling model. Zou et al. (2023) explored LLM attacks but focused on inducing objectionable generated content from LLM APIs. Similarly, You, Hammoudeh, and Lowd (2023) used GPT-3.5 API to introduce diverse style-based triggers into sentences. While these attacks represent an intriguing category of adversarial methods, they are significantly less explored compared with character-level and word-level attacks, and, in general, do not suit our scenarios of interest. (For example, a review to be posted to a reviewing Web site with a style changed from social media to biblical would not suit here; similarly, a focus on objectionable content does not align with our goal.) Further, they are not available on the *TextAttack* platform and are generally not framed in the same combinatorial optimisation manner, so we do not explore them in this article.

Attacks Against Other Tasks. While the majority of attacks in NLP have been against classifiers since relatively early on, there have been attacks against other tasks: Machine Translation (Belinkov and Bisk 2018), Textual Entailment (Kang et al. 2018), Named Entity Recognition (Srinivasan and Vajjala 2023), Grammatical Error Correction (GEC) (Wang and Zheng 2020), and so on. Often attacks in this space are designed for the purposes of robust training or adversarial testing (Huang et al. 2020; Wang and Zheng 2020; Xiao et al. 2023; Aghababaeyan et al. 2023), and so do not require the combinatorial optimization framing generally necessary for the classification task (Lei et al. 2019)—for example, Wang and Zheng (2020) produce adversarial GEC examples just by identifying "vulnerable" tokens and replacing them with grammatical errors people may make. Although interesting, these kinds of adversarial examples are out of the scope for this article.

2.2 Adversarial Text Quality

The primary goal of an adversarial attack is to change the model's prediction. Therefore, the main metrics for assessing adversarial text effectiveness revolve around the success of the attack (attack success rate) and the rate of perturbation. However, it would obviously be possible to fool the target classifier by radically changing the text, potentially to be unrecognizable. Therefore, adversarial text quality is also considered, in terms of properties like how similar the adversarial text is to the original, how grammatical it is, and so on. Generation methods will also include these considerations, often in the form of constraints, in their methods for producing adversarial texts. Both evaluation and constraints often use automatic metrics to measure these properties.

Properties Measured by Automatic Metrics. Morris et al. (2020a) identified four essential types of properties for evaluating adversarial texts: semantics, grammaticality, character-level overlap with the original texts, and non-suspicion towards human readers. We give some examples of how these properties have been operationalized.

For semantic similarity, adversarial texts are assessed on how well they preserve the semantics of the original texts. TEXTFOOLER (Jin et al. 2020), BAE (Garg and Ramakrishnan 2020), CLARE (Li et al. 2021), HQA-Attack (Liu et al. 2023), and GLARE (Chi et al. 2022) assess semantic similarity using Universal Sentence Encoder (USE)-based cosine similarity. In contrast, A2T (Yoo and Qi 2021) utilizes a fine-tuned DistilBERT model for similar assessments.

For assessing grammaticality of the adversarial texts, Pruthi, Dhingra, and Lipton (2019) deploy a word recognition model to evaluate character-level (misspelling) attacks, while TEXTFOOLER (Jin et al. 2020), HotFlip (Ebrahimi et al. 2018), ensemble transfer attack (Kwon and Lee 2022), and BAE (Garg and Ramakrishnan 2020) (in the BAE-R setting) replace words only with those of the same part of speech. Morpheus (Yildiz and Tantuğ 2019) enforces grammaticality with an inflection-constrained word swap.

Overlap properties quantify differences between original and adversarial texts. These can be carried out at the character level through methods such as Levenshtein edit distance (Gao et al. 2018) and *n*-gram-based measures like BLEU (Wang et al. 2020; Yildiz and Tantuğ 2019), as well as word-embedding distances (Jin et al. 2020; Wang et al. 2021b).

The fourth property type of Morris et al. (2020a), non-suspicion, has broadly speaking not yet been operationalized, and is discussed below.

Human Evaluation of the Adversarial Texts. Some algorithms are only evaluated by automatic metrics, as described above. However, others include a variety of human evaluations to verify the results of the automated evaluation. Common assessments include the validity of the adversarial perturbation, the classification accuracy of the original task by humans (Jin et al. 2020; Alzantot et al. 2018; Garg and Ramakrishnan 2020; Li et al. 2020), the similarity of the adversarial text 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).

A few works go beyond this to also ask about the naturalness of adversarial sentences. BAE (Garg and Ramakrishnan 2020) evaluated this by asking 3 annotators to judge whether each of the original sentences and three adversarial variants is from the original distribution, on a 1–5 Likert scale over 100 samples. PWWS (Ren et al. 2019) measured detectability by having human readers compare original-adversarial sentence pairs.

Human Perception of Suspiciousness. The aim of Morris et al. (2020a) was to propose a unified framework for evaluating adversarial texts, motivated in part by difficulty in fairly comparing some previous works. As noted above, one of the properties they proposed for evaluation was "non-suspicion," to be evaluated in addition to semantics, grammaticality, and overlap. For the two systems they were evaluating, TEXTFOOLER and GENETICATTACK (Alzantot et al. 2018), in addition to asking human judges about semantic preservation (measured by a 1–5 Likert scale) and grammaticality (measured by asking judges to spot any grammatical errors), they asked about suspiciousness, as follows.

To assess suspiciousness, judges had to determine whether each sentence was real or computer-altered (i.e., a binary judgment) among an even mix of real and adversarial sentences. A sentence was deemed suspicious if the percentage of judges identifying perturbed sentences as computer-altered exceeds a predetermined threshold. Morris et al. (2020a) only obtained judgments for TEXTFOOLER on their sentence-length movie review dataset, as it can be a time-consuming task; for each text, they obtained 10 judgments. Under this setup, the judges correctly nominated 69.2% as suspicious, against the baseline of 50% if adversarial and real were indistinguishable.

Table 2 shows a few example movie reviews and their human vs computer-altered annotations by 10 human judges. Among these examples, for sentences #1–#5, the original texts typically had more judges pick them as human (although not universally so), while the TEXTFOOLER texts varied quite widely in how many judges picked them as human-written. However, in the case of sentence #6, both the original and the adversarial sentences were chosen as human-written sentences 50% of the time.

Table 2
Pairs of sentences written by humans and adversarial variants of those sentences produced by TEXTFOOLER, from the Morris et al. (2020a) dataset. Hum/Com represents votes for human vs. computer-altered out of the ten binary votes per sentence. Sus. score is the proportion of votes for computer-altered out of the total vote.

Sentences	Hum./Com. (vote)	Sus. score
orig #1—the film is hampered by its predictable plot and paper thin supporting characters	9/1	0.1
adv #1—the film is restrained by its predictable conspiring and paper lean favours qualities	3/7	0.7
orig #2—davis is so enamored of her own creation that she can’t see how insufferable the character is	9/1	0.1
adv #2—davis is even enamored of her own creation that she can’t suppose how intolerable the character is	5/5	0.5
orig #3—flat out amusing, sometimes endearing and often fabulous, with a solid cast, noteworthy characters, delicious dialogue and a wide supply of effective sight gags	8/2	0.2
adv #3—flat out amusing, sometimes endearing and frequently mega, with a dependable cast, prodigious roles, luscious dialogue and a wholesale affords of actionable viewpoint gags	4/6	0.6
orig #4—payne has created a beautiful canvas, and nicholson proves once again that he’s the best brush in the business	10/0	0.0
adv #4—payne has fabricated a sumptuous canvas, and nicholson testifies once again that he’s the best brush in the business	4/6	0.6
orig #5—brian tufano’s handsome widescreen photography and paul grabowsky’s excellent music turn this fairly parochial melodrama into something really rather special	9/1	0.1
adv #5—brian tufano’s handsome widescreen photography and paul grabowsky’s excellent music turn this fairly parochial melodrama into something really rather singular	9/1	0.1
orig #6—a formula family tearjerker told with a heavy irish brogue accentuating, rather than muting, the plot’s saccharine thrust	5/5	0.5
adv #6—a formula families tearjerker told with a tremendous irish brogue accentuating, rather than muting, the plots’s moralizing thrust	5/5	0.5

Possibly, the presence of the unusual words "tearjerker," "brogue," and "accentuating" raises the suspicion of the readers. A wide range of levels of doubt about whether texts are human-authored can thus be observed in the Morris et al. (2020a) data.

Dyrmishi, Ghamizi, and Cordy (2023) expanded this evaluation by conducting an extensive human survey of 9 word-based attacks across 3 datasets, focusing on the validity and naturalness of the adversarial examples. They evaluated suspiciousness (referring back to the work of Morris et al. [2020a] here, although defining it specifically as "the percentage of adversarial texts recognised as 'computer altered"'), detectability, grammaticality, and meaningfulness as components of naturalness.

Their methodology involved a three-part questionnaire where an adversarial text was defined as "a text automatically altered by a program by replacing some words with others." In the first section, judges evaluated individual texts on validity, suspiciousness, detectability (with options to indicate if the text seems human-written or computer-altered), grammaticality, and meaningfulness, over five original and five adversarial texts without exposure to multiple versions of the same text. In the second section, to assess detectability, judges were shown four adversarial texts and given a number *n* indicating the count of altered words, with the task of identifying *n* modified words in the texts. The final section presented judges with both the original and adversarial versions of the same four texts they saw in the second part to assess grammatical error introduction and meaning preservation.

In terms of suspiciousness, they found that 60.33% of the adversarial examples were perceived as computer-altered overall, which significantly exceeded the 21.43% suspicion rate for original texts. Notably, five out of the nine adversarial attacks produce more than 25% of examples that are likely to be mislabelled by humans, indicating these examples would fail to achieve their intended effect in NLP systems checked by humans. Additionally, humans could detect nearly half (45.28%) of the perturbed words in the adversarial texts.

Both of these works have treated suspiciousness as binary, although it is clear that there are gradations of suspiciousness both across individual examples and in the aggregate. We use this as our starting point.

3. Creating the Human Suspicion Dataset

In this section, then, we describe the gathering and analysis of a dataset of adversarial examples annotated with graded judgments about suspiciousness. For these graded judgments, we use a 1–5 Likert scale, as Garg and Ramakrishnan (2020) did to evaluate what they described as "naturalness." Likert scales like this are used widely in NLP, both for human judgments related to adversarial texts—Jin et al. (2020) (TEXTFOOLER) and Garg and Ramakrishnan (2020) (BAE) use 1–5 scales for grammaticality checks, and Alzantot et al. (2018) assesses similarity to original texts on a 1–4 scale—and also for standard datasets across a variety of tasks that require graded human judgments, such as the Semantic Textual Similarity (STS) dataset of Cer et al. (2017) that is part of the GLUE benchmark (Wang et al. 2018).

We give natural language descriptions to our 5 categories to reflect levels of certainty about whether a text is one produced by a human (with a score of 1 being most human-like) or whether altered by a computer (a score of 5 being most likely to have been altered by a computer). These are presented in Table 3.

In the subsections below, we describe the data and methods we use to construct our adversarial examples and the methods for annotating them. We then present some descriptive statistics and analyses of the collected data.

Table 3
Likert-scale suspicion labels for human annotations.

Label	Description
1	Very confident that written by a human with no computer alteration
2	Probably written by a human with no computer alteration
3	Uncertain whether original or altered by a computer
4	Probably altered by a computer
5	Very confident that it is altered by a computer

3.1 Data Gathering Setup

Base Dataset. Following Morris et al. (2020a), we also use the Rotten Tomatoes Movie Review (MOVIEREVIEW) sentiment analysis dataset (Pang and Lee 2005) that is widely used in the adversarial example generation literature as well (Pruthi, Dhingra, and Lipton 2019; Alzantot et al. 2018; Ribeiro, Singh, and Guestrin 2018; Ren et al. 2019; Iyyer et al. 2018; Yoo and Qi 2021; Li et al. 2020, 2021; Jin et al. 2020). MOVIEREVIEW is a collection of 10,662 reviews gathered from the Rotten Tomatoes movie review database.[5] We split the dataset into 5,662 train and 5,000 test sets. It has 21 words on average and 269 words maximum per review. In all experiments, we use 128 maximum sequence lengths for the language models on MOVIEREVIEW.

We produce our adversarial texts using four extensively used attacks: PRUTHI (Pruthi, Dhingra, and Lipton 2019), ALZANTOT (Alzantot et al. 2018), TEXTFOOLER (Jin et al. 2020), and BAE (Garg and Ramakrishnan 2020), via their *TextAttack* library (Morris et al. 2020b) implementations. Details of the attack methods are as follows.

- PRUTHI (Pruthi, Dhingra, and Lipton 2019). This character-level attack modifies texts by randomly switching, deleting, adding characters, or introducing keyboard mistakes. In our experiments, we allow a maximum of half the words from the original text to be perturbed.
- ALZANTOT (Alzantot et al. 2018). This word-level attack permits the alteration of nearly every word in a sentence using context-preserving synonyms. The synonym search is done over a large search space including GloVe (Pennington, Socher, and Manning 2014) and counter-fitting word vectors (Mrkšić et al. 2016), and the Google 1 billion words language model (Chelba et al. 2014). We follow Jia et al.'s (2019) method for accelerated synonym search using a faster Transformer-XL model and limiting semantic drift by pre-selecting word neighbors. We also set the minimum perturbation limit to one-fifth of the maximum sequence length.
- TEXTFOOLER (Jin et al. 2020). This is a word-level attack that ranks the words by their impact on sentence predictions, and then sources synonyms using Counter-Fitted Word Vectors (Mrkšić et al. 2016). The

---

[5] `https://datarepository.wolframcloud.com/resources/Sample-Data-Movie-Review-Sentence-Polarity`.

method crafts a semantically equivalent sentence, controlled by the cosine similarity of their USE encodings (Cer et al. 2018), and iterates this process until it can alter the target model's prediction.

- BAE (Garg and Ramakrishnan 2020).This word-level attack leverages a pre-trained BERT MLM to create adversarial sentences by masking tokens and substituting them with contextually relevant synonyms. They argue that token-level similarity-based attacks such as TEXTFOOLER only compare word-embedding level similarity, not the overall text semantics and that this may lead to unnatural context-drifting. BAE enforces semantic similarity using the highest USE score (Cer et al. 2018) and a parts-of-speech filter for the synonym replacement. This method experiments on both token replacement (R) and insertion (I) techniques. We use the *TextAttack* (Morris et al. 2020b) library's BAE-R (token replacement) setting.

Target Model. For generating adversarial texts, we use a pre-trained BERT-base-cased model, adding a fully connected dense layer of 768 nodes, a layer of 50% dropout, and another dense layer of 768 nodes following Tonni and Dras (2023). The dataset split is 80-20 train-test. We train the model for 3 epochs with $5e^{-5}$ learning rate and AdamW optimization without freezing any layer of the backbone model. This $\text{BERT}_{\text{BASE}}$ model achieves 90.94% test accuracy on the MOVIEREVIEW dataset on sentiment classification task. The accuracies of the clean model and the model under attack are given in Table 4. We note that in all cases, TEXTFOOLER degrades the classifier the most, followed by BAE.

Table 4 also lists the number of correctly predicted test records and the generated adversarial texts from them. The BERT classifier predicts correct labels for 4,350 out of 5,000 test sentences. The adversarial sentences are generated on this correctly predicted test set. Among all the attacks TEXTFOOLER is most successful in attacking the target model by reducing BERT's accuracy to only 16%.

Amazon Mechanical Turk (MTurk) Setup. Our primary annotations are collected through Amazon MTurk.

We selected 540 original sentences and their four adversarial versions (in total 5 sets), giving 2,770 sentences altogether; these were to have one human judgment each. Further, we sampled 63 additional sentences from the 5 sets ($63 \times 5 = 315$ sentences); these were to have three human judgments each, giving 945 ($= 3 \times 315$) annotations.

Table 4
Number of correctly predicted original test sentences and generated adversarial sentences by $\text{BERT}_{\text{BASE}}$ classifier and its accuracy under attack.

	No. of sentences	Test accuracy
Original	4,350	0.909
PRUTHI	3,023	0.518
ALZANTOT	2,999	0.430
TEXTFOOLER	4,200	0.160
BAE	3,003	0.313

Table 5
Frequency of each Likert-scale suspicion score [1 . . . 5] for original human and adversarial sentences provided to the human annotators, along with mean suspicion scores. "Adversarial" represents the total of all adversarial methods.

Sus. score	1	2	3	4	5	Mean score
Total	944	1,062	254	521	234	2.35
Original	234	227	39	83	20	2.05
Adversarial	710	835	215	438	214	2.42
PRUTHI	147	217	58	124	57	2.55
ALZANTOT	203	225	45	91	39	2.23
TEXTFOOLER	210	209	56	88	40	2.24
BAE	150	184	56	135	78	2.68

(We refer to this latter as the "common set," and its purpose is to enable an assessment of the level of agreement.) Altogether we collected 3,715 annotations. The Turkers were given instructions derived from that Morris et al. (2020a):

> All the sentences are based on texts originally written by humans. However, some of them have been altered (in terms of changes in characters, words, etc.) by a computer algorithm aiming to produce texts that can fool another computer system in some way.

Sentences were grouped into batches of 53 sentences (50 sentences for our annotation purposes and 3 sentences for attention-checking purposes) and Turkers selected a batch.[6] (The attention-checking sentences are excluded in the result evaluation.) In each batch, there was an equal mix of sentences of each of the 5 types (original sentences plus 4 adversarial types of sentences). The batch was created in such a way that there was no repetition of the different variants of one sentence, and the Turkers would not see the same sentence twice. The sentences within a batch were presented one at a time. When presented with a sentence, the Turker was asked to assign a score of 1–5 to the text based on the descriptions in Table 3.

The reward per sentence (HIT) was 0.13 USD (equivalent to 0.20 AUD). We chose this cost based on some small-scale informal surveys among colleagues that estimated the time allocation for completing the task was one minute per text; this would give a payment rate that meets the Australian minimum wage. Surveyors were master-level Turkers with a HIT approval rate of 95%. The ethics approval letter was disclosed and linked with the survey.[7] The MTurk survey interface is illustrated in Appendix Figure A.1.

3.2 Analyses

Distribution of Annotation Scores. The distribution of annotation scores for all texts is given in Table 5 (raw counts) and Figure 1 (percentages). Across both original texts and all adversarial texts, scores of 1 or 2 (very confident or probably written by a human)

---

[6] There was no restriction on the number of batches each Turker could accept.

[7] Ethics approval by Macquarie University (Human Ethics Committee Approval Code: 5201800393).

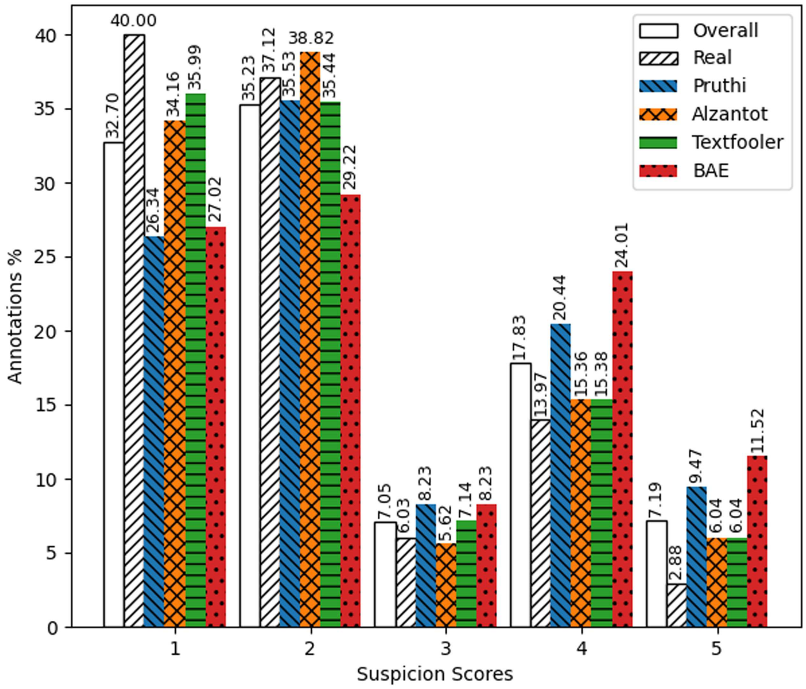


Figure 1
Proportion (%) of human suspiciousness annotations grouped by the levels of the Likert scale labels [1 . . . 5].

are the most common scores, with 4 (probably altered by a computer) the next most common. The score 5 was relatively rarely chosen, indicating that annotators found indications of computer alteration to be relatively subtle. The score of 3 was likewise relatively rarely chosen, indicating that annotators did want to avoid being undecided.

This distribution of the mean scores for all text types were between 2 and 3. As expected, original texts have the lowest (most human-like) scores, but the adversarial text methods all fall on the human-like side of undecided as well. The means are significantly different using a one-way ANOVA, and a post-hoc Tukey test reveals that they fall into two groups: the human, ALZANTOT, and TEXTFOOLER (with the two adversarial methods falling just outside being differentiable from human at $0.06 < p < 0.07$) as the first group; and PRUTHI and BAE as the second group.

Subsequently, we examine the reliability of our dataset in several ways.

Annotator Agreement. In light of work that shows that humans are not very good at recognizing machine-generated text (Karpinska, Akoury, and Iyyer 2021; Clark et al. 2021), including adversarial examples (Dyrmishi, Ghamizi, and Cordy 2023), we look to see whether there is some level of agreement about suspicion, or whether it is essentially random. While there is a degree of common practice within the NLP community about metrics for evaluating agreement on labels representing a *nominal* factor (Cohen's kappa, Krippendorff's alpha, etc.) there is no similar degree of common practice for *ordinal* factors like ours. We therefore use the metric of Vogel, Koutsombogera, and Costello (2020), which quantifies inter-annotator agreement in Likert scale labels by calculating the average absolute deviation from the median response for each item. For

C annotators and k Likert categories, normalized annotator disagreement of each item i is:

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

where $\tilde{k}_i$ is the median response to item i and k_i^c is the category supplied by annotator c to item i.

Normalizing the δ_i by the number of items n gives the average annotator disagreement level:

$$\delta = \frac{\sum_{i=1}^{n} \delta_i}{n} \tag{2}$$

For our Likert scale of 5 points with $C = 3$ annotators, the smallest possible value for δ_i is 0 (perfect agreement) and the largest possible value is 1.33. Consider the unnormalized scores for each item $C\delta_i$: These can come from the set $\{0, 1, 2, 3, 4\}$, with the largest disagreement of 4 occurring when one annotator chooses label 1 and another label 5; δ_i in this case is 1.33. If all items were annotated such that their $C\delta_i$ were 4, δ would likewise be 1.33.

We calculate our annotator agreement score δ over the 315 sentences receiving three judgments. Table 6 presents the unnormalized per-item disagreement frequency $C\delta_i$ and the overall disagreement level δ. Of the 315 sentences, unanimity ($C\delta_i = 0$) was found in only 43 cases. Similar results are observed for both original and adversarial sentences. The overall δ is 0.61, a little under halfway between perfect agreement and maximal disagreement. This level of label variability aligns with humans finding this task difficult but there still is some shared intuition about the suspiciousness of texts.

The lowest level of disagreement is within the original human-authored sentences. Within the adversarial examples, sentences altered by the BAE word-level attack ($\delta = 0.71$) caused the most disagreement, followed by the PRUTHI character-level attack ($\delta = 0.64$). These attacks are also the ones that have the highest mean scores in Table 8, suggesting that these methods may generate some instances that contain perturbations that are more unusual and noticeable to some subjects, whereas the other methods are more consistent.

Non-MTurk Annotators. In light of Karpinska, Akoury, and Iyyer (2021) on poor performance of Turkers evaluating generated text (albeit open-ended), as a check on our Turker annotations we also recruited three colleagues who are native speakers of

Table 6
Average inter-annotator disagreements δ (Eqn (2)) the frequency of $C\delta_i$ (Eqn (1)) obtained from MTurk annotators. Higher scores mean more disagreement.

	Overall	Original	PRUTHI	ALZANTOT	TEXTFOOLER	BAE
δ	0.61	0.50	0.64	0.59	0.61	0.71
$C\delta_i = 0$	43	13	7	11	6	6
$C\delta_i = 1$	104	24	20	24	26	10
$C\delta_i = 2$	57	9	12	4	9	23
$C\delta_i = 3$	88	15	19	17	20	17
$C\delta_i = 4$	23	2	5	7	2	7

Table 7
Average inter-annotator disagreements δ (Equation (2)) the frequency of $C\delta_i$ (Equation (1)) obtained from non-MTurk annotators. Higher scores mean more disagreement.

	Overall	Original	Adversarial
δ	0.65	0.62	0.66
$C\delta_i = 0$	4	1	3
$C\delta_i = 1$	20	4	16
$C\delta_i = 2$	31	7	24
$C\delta_i = 3$	16	2	14
$C\delta_i = 4$	4	1	3

English, and not part of this work and unaware of the purpose of the study, to annotate a sample of 75 sentences (5 × 15 sets).

Table 7 shows that the agreement levels for this smaller sample mirrors Table 6, except that the adversarial examples are all grouped together. The overall δ of 0.65 is very similar to the MTurkers' 0.61 (and in fact, even representing slightly greater disagreement than the MTurker results), giving support to the results of Table 6 being representative of the task. Also as with the MTurkers, there was better agreement (here only slightly) on original texts than on adversarial ones.

Links to Detection of Adversarial Examples. As another reliability check, we relate our data to the prior work of Dyrmishi, Ghamizi, and Cordy (2023), where their suspiciousness regarding whether a text is human-generated or computer-altered is measured by a binary question. An obvious possible transformation of our data to binary is to consider scores of 1 and 2 as non-suspicious or human-like and scores of 4 and 5 as suspicious or computer-generated; this reflects the textual instructions given to our annotators, and we refer to this as the symmetry-split. An alternative transformation, the 1-vs-other split, considers only the score of 1 as non-suspicious and all other scores as indicative of some degree of suspicion. Table 8 displays results from both types of splits. In the symmetry-split scenario, we would interpret this as saying that annotators only recognize 27% of all adversarial texts as being computer-altered, which is ridiculously low. However, in the 1-vs-other scenario, the percentage of texts raising suspicion increases

Table 8
Proportions of suspiciousness scores split in symmetry (scores of 1,2 human-authored and 4,5 computer-altered) and 1-vs-other (score 1 human-authored and 2–5 computer-altered). (Note that proportions under symmetry split do not add up to 1 as the score of 3 is excluded.)

	Symmetry split		1-vs-other split	
	Human (1,2)	Computer (4,5)	Human (1)	Computer (2–5)
Overall	0.67	0.25	0.31	0.69
Real	0.76	0.17	0.39	0.61
Adv	0.64	0.27	0.29	0.71
PRUTHI	0.60	0.30	0.24	0.76
ALZANTOT	0.71	0.21	0.34	0.66
TEXTFOOLER	0.69	0.21	0.35	0.65
BAE	0.55	0.35	0.25	0.75

to 69%, aligning more closely with the findings of Dyrmishi, Ghamizi, and Cordy (2023) (60.3%).

In terms of the rankings of adversarial methods by suspiciousness, we see a small difference in this 1-vs-other binarization versus the mean Likert score of Table 5. Under the mean Likert scores, BAE is the worst, followed by PRUTHI (although they are not statistically significantly different), whereas this is reversed in 1-vs-other. This is due to BAE having the largest number of 5 scores (most obviously computer-altered) in Likert counts by some margin, but it also does slightly better than PRUTHI on 1 score. This more extreme variability of performance by BAE is to some extent concealed by the binarization. ALZANTOT and TEXTFOOLER are closer to each other, and closer to humans than the other two, in both rankings.

Our rankings have some differences from Dyrmishi, Ghamizi, and Cordy (2023). There, as here, ALZANTOT was observed to generate the most human-like sentences—only 46.5% were suspected to be computer-altered there—but the ordering of BAE (50.6%) and TEXTFOOLER (61.5%) are the reverse of ours. We are not sure what factors might lead to this.

Impact of the Survey Instructions: Prevalence. A known effect in psychology is that humans can make judgments about items based on their belief about prevalence (Levari et al. 2018, on the "blue dot effect"). In our primary data-gathering setup, no information about prevalence is provided; this is reflective of our scenarios of interest (as, for example, when a person sees a single review with typos and speculates about whether it was authentically written by a human or modified). However, we may also wonder how scores would be affected if human judges were provided with information about prevalence, and if this would change overall observed patterns. In our setup, where annotators are not provided with information about prevalence, they may expect an even split between human and computer-altered sentences (which is also the usual setup in the work described above). Such an assumption about prevalence might incline annotators towards identifying more adversarial sentences as real. To investigate whether this is the case, and see whether providing information about prevalence changes the patterns we have observed, we conducted a smaller survey with 750 sentences with the same distribution of original and adversarial sentences, consisting of 510 single annotations across 102 × 5 sets of sentences, and 3 annotations on 16 × 5 sets of sentences, explicitly instructing the participants that there were more computer-altered sentences than real ones. The results, presented in Table 9, indicate that this change in instruction does not

Table 9
Frequency of each Likert-scale suspicion score [1 . . . 5] for original human and adversarial sentences provided to the human annotators, along with mean suspicion scores, under the prevalence-informed setting. "Adversarial" represents the total of all adversarial methods.

Sus. score	1	2	3	4	5	Mean score
Total	143	296	84	176	51	2.60
Original	38	60	14	30	9	2.42
Adversarial	105	236	70	146	42	2.64
PRUTHI	25	65	22	40	13	2.70
ALZANTOT	29	63	19	33	6	2.49
TEXTFOOLER	30	56	19	34	10	2.58
BAE	21	52	10	39	13	2.79

Table 10
Mean metric score between pairs of original and adversarial sentences generated by each method. "Overall" represents scores across all methods. r is the Pearson correlation coefficient with respect to human suspicion scores.

Metrics	PRUTHI	ALZANTOT	TEXTFOOLER	BAE	Overall	r
BLEU	0.869	0.752	0.827	0.838	0.856	−0.057
METEOR	0.947	0.908	0.935	0.935	0.945	−0.064
ROUGE1	0.941	0.882	0.922	0.928	0.934	−0.064
ROUGE2	0.893	0.794	0.859	0.869	0.882	−0.060
ROUGE3	0.844	0.713	0.798	0.810	0.831	−0.058
ROUGEL	0.941	0.882	0.922	0.928	0.934	−0.064
$\text{BERTScore}_{\text{F1}}$	0.985	0.985	0.986	0.984	0.985	−0.094
$\text{BERTScore}_{\text{Prec}}$	0.981	0.984	0.985	0.983	0.983	−0.106
$\text{BERTScore}_{\text{Rec}}$	0.989	0.986	0.988	0.985	0.987	−0.074
MeaningBERT	0.877	0.878	0.877	0.875	0.877	0.023
BLEURT	0.126	0.489	0.448	0.159	0.428	−0.138

significantly alter the outcomes compared to those shown in Table 5. The annotators seem slightly more suspicious overall—the mean score across all texts is 2.60 versus 2.35 under the main setting—but the relative rankings are all the same, with original sentences scoring lowest (most human-like), PRUTHI and BAE the most suspicious, and the other two methods in the middle. Other patterns from the original setting likewise show themselves (e.g., scores of 3 and 5 chosen infrequently). As well as indicating that instructions regarding prevalence do not matter particularly for assigning suspiciousness scores—although they might for human detection of computer-altered text—this second albeit smaller dataset supports the conclusions of the original.[8]

Correlation with Text Evaluation Metrics. As we observed in Section 2, adversarial example generation methods generally implement some constraints on various properties to improve example quality, which include semantic and other similarity to the original, measured by a range of metrics. However, these methods do not consider the non-suspicion as a novel characteristic to be measured. We therefore examine the extent to which our non-suspicion scores are in fact distinct from the kinds of metrics used to measure overlap with the original sentence. Table 10 presents the average scores for BLEU (Papineni et al. 2002), METEOR (Banerjee and Lavie 2005), ROUGE (Lin 2004), BERTScore (Zhang, Kishore, et al. 2020), MeaningBERT (Beauchemin, Saggion, and Khoury 2023), and BLEURT (Sellam, Das, and Parikh 2020) of the adversarial sentences, alongside their correlation with human suspicion ratings. We see that the non-suspicion scores have almost no relationship to other metrics, indicating it is in fact a different dimension of adversarial example quality.

---

[8] For completeness, we also looked to see if telling annotators that there were fewer adversarial texts than original human ones affected the scores. This differs from our main prevalence experiment in that we are there, on data containing original to adversarial sentences in the ratio 1:4, contrasting the cases of no information about prevalence with true information about prevalence. Here, using data with the same distribution means that we give them *incorrect* information about prevalence. Consistently with the earlier results, annotators do take prevalence information into account when it is available (even if false): They are slightly less suspicious overall as measured by mean score, but with other patterns—such as original sentences' mean scores being lower than all adversarial sentence mean scores—remaining the same.

4. Predicting the Level of Suspiciousness

In this section, we look to build a predictor of the scores described in Section 3. In doing this, we look at other regression tasks based on subjective opinions such as the evaluation of abstractness through human imperceptibility in SemEval 2021 (Zheng et al. 2021) and the assessment of semantic textual similarity in SemEval 2017 (Cer et al. 2017). The SemEval 2021 task involves rating words based on their abstractness, while the SemEval 2017 task focuses on evaluating multilingual and cross-lingual sentences' semantic similarity on a 0-5 Likert scale. We thence build several regressors that are trained using scores from human annotations to predict human suspicion scores on adversarial texts.

4.1 Experimental Setup

Likert Score Dataset. Our primary dataset for our prediction models is the dataset described in Section 3, consisting of 3,015 total annotated sentences from the MOVIEREVIEW dataset. From this, we take the 315 sentences with 3 annotations each as the test set and use the median as the assigned Likert score for the item. Of the remaining 2,707 (single annotation) sentences, we split them 90%–10% as train and dev.

Morris Dataset. Morris et al. (2020a), as noted above, collected binary suspicion judgments on a small set of 100 unique sentences from the MOVIEREVIEW dataset and their TEXTFOOLER generated version (so 200 texts in total). As each of these was judged by 10 annotators, we can derive a scalar score from this by taking the average of the annotations, treating human as 0 and computer-altered as 1; the score is thus between 0 (all annotators think computer-altered) and 1 (all annotators think human), with a score of 0.5 indicating most uncertainty. We use 144, 20, and 36 sentences as the train, dev, and test set. This dataset, due to its small size, is very much secondary and only used as a check on our Likert dataset. Table 2, with its illustrations of a few sentences from the Morris dataset, includes our derived suspicion score in the rightmost column.[9]

Evaluation Metrics. Like Cer et al. (2017), we consider Pearson's correlation to be our primary evaluation metric, but we also report Spearman's rank correlation and root-mean-squared-error (RMSE) scores.[10]

4.2 Regressor Models

The core of our regression models is a language model over the texts. However, following the best regression models in Semantic Textual Similarity SemEval task (Cer et al. 2017), we use additional numerical features. We then combine these together in two ways: (1) **TEXT+NUM**, combining the prediction of two regressors separately trained on texts and numerical features by averaging their predictions; and (2) **MULTIMODAL**, training multimodal regressors on both text and numerical features using multimodal Transformers as proposed by Gu and Budhkar (2021).

---

[9] It is not possible to create a similar dataset from Dyrmishi, Ghamizi, and Cordy (2023), as almost all of their sentences have only a single binary judgment.

[10] Multiple metric perspectives can be useful. Consider Pearson's r and Spearman's ρ (van den Heuvel and Zhan 2022): The former is considered a measure of linear association, while the latter is considered a measure of monotonic association. While they give similar scores on normally distributed data, on non-linear but monotonic data ρ scores are systematically higher. In addition, ρ is less sensitive to outliers; consequently, if there are major outliers, ρ will differ from r.

Language Model Component. For these, we fine-tune BERT$_{\text{BASE}}$, RoBERTa, XLNET, DistilBERT, and ALBERT. In addition, we fine-tune BERT and RoBERTa models using the SMART approach of Jiang et al. (2020). SMART incorporates smoothness-inducing regularization and Bregman proximal point optimization techniques, preventing overly aggressive updates and ensuring the model remains within a small neighborhood of its previous iteration at each step. SMART achieves the highest performance on the STS semantic similarity benchmark (Cer et al. 2017)[11] and is top-performing in the Adversarial GLUE multi-tasking benchmark (Wang et al. 2021a) for evaluating language model robustness. We refer to these models as SMART$_{\text{BERT}}$ and SMART$_{\text{RoBERTa}}$.

All the regressors are trained from 1 to 10 epochs with AdamW optimizer 0.01 weight decay. We choose the regressors having the maximum Pearson's correlation on the devset. For the Morris dataset, just consider BERT, RoBERTa, SMART$_{\text{BERT}}$, and SMART$_{\text{RoBERTa}}$ regressors for comparison.

4.3 Numerical Features

We use the features below and train several models on them for use in the **TEXT+NUM** setup: LinearRegressor, HuberRegressor, RandomForestRegressor, GradientBoostingRegressor, XGBoostRegressor, VotingRegressor, and StackingRegressor. In these models, all numerical features are normalized using quantile normal transformation.

- ***Rate of perturbations.*** This feature is represented by the percentage of words modified by the adversarial attack on a sentence. **(1 feature)**
- ***Word embeddings.*** We use GoogleNews word2vec (Mikolov et al. 2013) vectors, with 300 dimensions. We represent each sentence by concatenating the average pooled vector representations of its words. Additionally, each word vector is weighted by its Inverse Document Frequency value. We take each element of the vector as a feature. **(300 features)**
- ***Adversarial text detectors.*** Automated adversarial text detection has shown notable success (Wang, Xiong, and He 2022; Mosca et al. 2022; Bao, Wang, and Zhao 2021; Nguyen-Son et al. 2023; Tonni and Dras 2023). Several of these methods use learned representations to identify properties that distinguish human-authored texts from adversarial ones. Incorporating features from these detectors may enhance our suspicion regressor's capabilities. From Tonni and Dras (2023), we utilize the influence function (IF) score, and DKNN-based useful and harmful ranks and distances, as used in the effective NNIF detector[12] **(5 features)**. Additionally, from Liu, Dras, and Zhang (2022), we calculate the local intrinsic dimensionality (LID) scores for each layer of BERT as well as the average LID. **(13 features)**

---

[11] `https://paperswithcode.com/sota/semantic-textual-similarity-on-sts-benchmark`.

[12] NNIF investigates a combined viewpoint of identifying the neighborhood subspace of the adversarial sentences based on influence function (Koh and Liang 2017) and DKNN (Papernot and McDaniel 2018), following Cohen, Sapiro, and Giryes (2020). Interested readers can refer to Tonni and Dras (2023).

- ***Grammaticality scores.*** Grammatical accuracy of these perturbations may correlate with human suspicions. We use both the GramFormer[13] and LanguageTool (adapted from *TextAttack*) packages. **(2 features)**
- ***GPT-4 produced scores.*** Chiang and Lee (2023) use LLMs to produce human-like scoring on texts and show that ChatGPT in particular chooses human written texts over adversarial texts in evaluating fluency and meaning preservation. Thus, besides collecting human annotations, we also design a GPT LLM (GPT-4 pointing to GPT-4-0613) using the OpenAI API, to predict scores of suspicion level on a scale of 1–5, and give a reason. The instruction contains three example sentences along with explanations of their scoring (Appendix Table B.1). Appendix Table B.2 presents a few examples of suspicion scoring and reasoning by ChatGPT. **(1 feature)**

Overall, this gives 322 numerical features.

4.4 Results and Discussion

Overall. Table 11 presents the Pearson and Spearman correlation coefficients, along with the RMSE scores. Overall, there is a moderate correlation (Pearson's $r = 0.515$ for the best model), demonstrating that this is a challenging task; we expect this is a reflection of the level of agreement among annotators discussed in Section 3. The Pearson and Spearman coefficients are close, so we do not see any evidence of major outliers, and so on.

Ensembling Separately Trained Text and Numerical Models Perform Better (Table 11). From Table 11, using only the texts, we achieve $r = 0.396$ with a BERT model and a closely related outcome of 0.377 with $\text{SMART}_{\text{BERT}}$. Among the ML regressors focusing solely on numerical features, RANDOMFOREST achieves the highest Pearson's r at 0.456. We enhance predictive accuracy by ensembling the predictions from these two models in a TEXT+NUM setting following Tian et al. (2017). The best correlation with $r = 0.515$ is achieved by combining predictions from the best TEXT+NUM model. Using the multimodal setting incorporating both texts and numerical features, in contrast, the highest Pearson's r achieved is 0.301.

Not All Numeric Features Perform Equally (Table 12). Additional experiments on the subsets of the numerical features show that combining some of the features with the texts gives better results than the others. We experiment with training the regressors using the NNIF features, grammaticality scores by GramFormer, and LanguageTool and GPT-4 generated scores paired with the TEXT; the results are illustrated in Table 12. There is a significant level of difference in the regressor outcome and adding the GPT scores (TEXT +GPT and TEXT +NNIF +GPT) with the text improves the result considerably.

GPT-4 as a Suspicion Scoring Baseline (Figure 2). Although GPT-generated scores work well as a training feature, the distribution of the suspicion levels (ranging from 1 to 5)

---

[13] `https://github.com/PrithivirajDamodaran/Gramformer/tree/main`.

Table 11
Pearson's and Spearman's correlation coefficients and the RMSE values between the various regressor scores and the human suspicion scores. (↑ indicates higher and ↓ indicates lower is better).

Regressor type	LM	Pearson's r ↑	Spearman's ρ ↑	RMSE ↓
TEXT	BERT	0.396	0.323	1.311
	RoBERTa	0.336	0.310	1.220
	AlBERT	0.119	0.123	2.242
	DistilBERT	0.192	0.198	2.613
	XLNet	0.074	0.071	2.580
	SMART$_{\text{BERT}}$	0.377	0.322	1.342
	SMART$_{\text{RoBERTa}}$	0.362	0.323	1.259
NUMERIC	VotingReg.	0.245	0.249	1.259
	StackingReg.	0.120	0.129	1.331
	GradientBoosting	0.195	0.198	0.234
	RandomForest	0.456	0.408	1.235
	XGBoost	0.112	0.133	1.350
	LinearReg.	0.214	0.218	1.322
	HuberReg.	0.183	0.189	1.402
TEXT+NUM	BERT+RandomForest	**0.515**	**0.585**	1.125
MULTIMODAL	BERT	0.301	0.303	1.219
	RoBERTa	0.272	0.269	1.244
	AlBERT	0.141	0.184	1.368
	DistilBERT	0.061	0.081	1.969
	XLNet	0.216	0.210	2.016
	SMART$_{\text{BERT}}$	0.274	0.274	1.264
	SMART$_{\text{RoBERTa}}$	0.173	0.196	1.318

Table 12
Pearson's *r* obtained by the MULTIMODAL regressors using text and different sets of features. Higher is better.

TEXT+	BERT	RoBERTa	SMART$_{\text{BERT}}$	SMART$_{\text{RoBERTa}}$
NNIF	0.272	0.282	0.287	**0.323**
NNIF +Gram.	0.289	0.288	0.289	0.322
NNIF +Lang.	0.272	0.272	0.244	0.307
GPT	0.351	**0.407**	**0.382**	0.221
GPT+Gram.	0.280	0.306	0.280	0.280
NNIF +GPT	**0.386**	0.218	0.335	0.233
NNIF +GPT+Gram.	0.320	0.242	0.316	0.264

determined by ChatGPT in Figure 2 suggests that it works better as an adversarial detector than in predicting human suspicion scores as such.

Regressors Trained on Subsets of the Data (Table 13). We also examined how well regressors performed that were trained on the data subset by type of text (original text and each adversarial method). The Pearson's correlation coefficients obtained from regressors trained just on BERT textual representations are presented in Table 13. Interestingly, the regressor trained on original text was best not only overall and specifically against other original text but also the best against PRUTHI (by a large margin) and TEXTFOOLER,

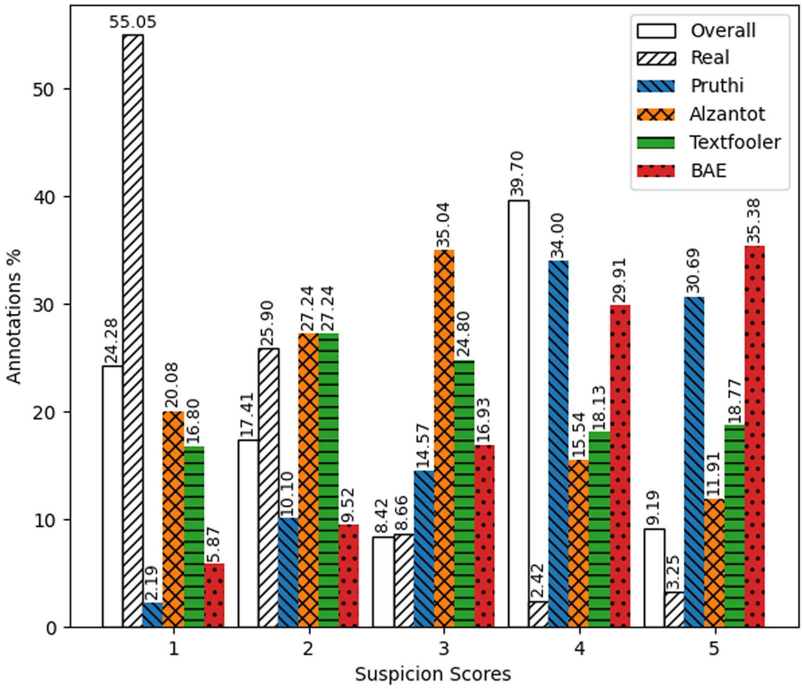


Figure 2
Distribution (%) of the ChatGPT (GPT-4) annotations grouped by the Likert scale levels.

Table 13
Pearson's *r* obtained by the regressors trained only on each subset of the train data and tested on all test data and the subsets. For comparison, the values when all data were used for training are also included. Higher is better.

	Test subset					
Train subset	Overall	ORIGINAL	PRUTHI	ALZANTOT	TEXTFOOLER	BAE
All	0.52	0.50	0.49	0.22	0.37	0.15
ORIGINAL	**0.30**	**0.45**	0.43	0.15	0.29	0.08
PRUTHI	0.22	0.08	0.05	0.20	0.24	**0.25**
ALZANTOT	0.16	0.08	0.07	0.16	0.02	0.01
TEXTFOOLER	0.15	0.07	0.11	**0.27**	0.04	0.13
BAE	0.17	0.31	0.30	0.20	**0.32**	0.07

even better than the regressors trained on that data. This suggests that better predictors of adversarial example suspiciousness could be built counter-intuitively by gathering more data only on original human text.

Results on the Morris Dataset. We achieved similar although slightly better correlations with human judgments using regressors built on the Morris dataset utilizing only textual data. Specifically, we achieved a Pearson's correlation coefficient *r* of 0.512, 0.506, 0.495, and 0.594 using BERT, RoBERTa, $\text{SMART}_{\text{BERT}}$, and $\text{SMART}_{\text{RoBERTa}}$, respectively. While not too much can be made of the results on such a small dataset, it is possible that collecting a large number of binary judgments and deriving scores from them would produce more systematically predictable scores. This would, however, be more expensive than collecting Likert scores.

5. Generating Low-suspicion Adversarial Texts

A possible application of the regressors presented in Section 4 is to use them to create adversarial texts that are less obvious to humans. Such an application would be another way of assessing the effectiveness of the regressors and the dataset used to create them. In this section, to illustrate this kind of application, we build an adversarial example generator that incorporates a non-suspicion constraint, with the aim that the examples generated will look less suspicious to humans. We compare the original adversarial examples against their suspicion-constrained (SUSCONSTRAINED) variants in a human evaluation and show that the suspicion-constrained ones are perceived as better.

5.1 Experimental Setup

Generator. We selected the TEXTFOOLER method in the *TextAttack* framework as our base generator, as it has the highest number of Likert score 1 ratings (i.e., most human-like) in our human evaluations (Table 5), as well as being most effective in fooling a classifier (Table 4).

We then produced a variant of this incorporating a suspicion constraint that requires the algorithm to generate adversarial sentences that maintain suspicion levels below a specified threshold; this variant thus has all of the constraints in TEXTFOOLER plus this additional one. Suspicion scores for each candidate sentence are determined using our best BERT-based TEXT+NUM regressor (Table 11).

There are many possible ways to set a threshold for the constraint—for example, as a function of the difference between the original text's regressor score and that of its TEXTFOOLER adversarial counterpart. We choose the simplest kind of threshold, a fixed value.

Selected Texts. We considered how many texts from the test set are available under two different thresholds: $\tau = 2.5$ and $\tau = 3.0$. We used $\tau = 2.5$ as our starting point in order to capture particularly bad TEXTFOOLER texts; this is above the mean Likert scores for original and TEXTFOOLER texts from Table 5. We then examined which texts could have their predicted suspicion score reduced. For $\tau = 2.5$, there are 61 TEXTFOOLER texts with regressor scores > 2.5, and 42 (80.3%) of those can have their suspicion score reduced, whereas a threshold of $\tau = 3.0$ results in only 27 TEXTFOOLER texts of which 20 (74.1%) can be reduced. This latter is too small, so we chose $\tau = 2.5$ for the experiment. The mean reduction in predicted suspicion score for this threshold was 0.63, with a standard deviation of 0.52.

Because we are using just a simple single fixed threshold for the maximum suspicion score, adding the constraint can drive the suspiciousness even below the suspiciousness score of the original text; this might flag an implausible score, as it is unlikely that TEXTFOOLER perturbation instances would make a sentence more human-like than the original human-authored one. On the other hand, we do not want to necessarily remove all whose scores are below the original, as this might result in too few sentences to test. Heuristically, we consider only those SUSCONSTRAINED variants whose scores fall below the corresponding original text's score by 0.2, which is half the gap between original and TEXTFOOLER text mean scores. (A more sophisticated constraint system could instead include, for example, a lower bound on non-suspicion that is a function of the original text.) This gave 23 texts.

MTurk Setup. Using the 23 selected texts, we conducted two subsequent surveys on Amazon MTurk to compare the quality of TEXTFOOLER and SUSCONSTRAINED texts. For the first, we structured the survey to present text pairs, each consisting of a TEXTFOOLER sentence and its suspicion-constrained counterpart, and asked the annotators to determine which sentence of the pair appeared to be written by a human: That is, it was a forced-choice preference testing setup. We gathered 5 judgments on each sentence pair and took the overall judgment on the majority vote. For the second survey, to see the effect on Likert score ratings for comparability with Section 3, we gathered the ratings in the same manner as in Section 3.1 (i.e., Turkers were presented with individual texts in random order and asked to rate on the provided scale, etc.).

5.2 Analysis

We consider first the forced-choice data. For each of the 23 pairs of texts, in 18 cases the SUSCONSTRAINED text was preferred and in 5 cases the TEXTFOOLER was preferred. Exact binomial tests are commonly used across many fields to test for significance in this kind of preference testing setup (O'Mahony 1986, for example); we use the usual $\alpha = 0.05$. The two-sided test result in this case is $p = 0.0106$, indicating that in our small sample, the SUSCONSTRAINED texts are significantly better.

For the Likert scale data, the TEXTFOOLER texts had a mean Likert score of 2.61[14] versus 2.26 for SUSCONSTRAINED texts, a difference in mean score of 0.35. For context, this improves our particularly bad TEXTFOOLER texts to score approximately the same as the overall TEXTFOOLER texts (mean Likert score 2.24) from Table 5; for an alternative perspective, it is approximately the same difference as between human-authored sentences (mean 2.05) and adversarial sentences overall (mean 2.42) in Table 5.

Table 14 gives some examples: It notes the vote among the 5 annotators for the SUSCONSTRAINED and plain TEXTFOOLER versions, the human Likert scores (for original and TEXTFOOLER) and predicted scores (for all three texts). In the first case, all annotators preferred the suspicion-constrained version, for which the suspiciousness score predicted by the regressor is much better than the TEXTFOOLER one; the strength of the vote (5–0) is not surprising, given how bad the TEXTFOOLER variant is compared to the original, both in human and predicted scores (1 vs 4, 1.76 vs 3.36), and the closeness of the SUSCONSTRAINED's predicted score to the original (1.76 vs 1.83). In the second case, the vote was marginal in favor of the SUSCONSTRAINED version. Here, the predicted suspiciousness score for the suspicion-constrained version is about halfway between the original and the TEXTFOOLER versions, a smaller relative improvement than in the first case (2.76 vs 2.46). In the third case, it is the TEXTFOOLER version that is preferred by a vote of 3 to 2. Here, the SUSCONSTRAINED version actually has a predicted score that is lower than the original (2.51 vs 2.38), which is an indication that the score is perhaps not likely to be accurate in this case; a more sophisticated constraint setup might have ruled this one out.

---

[14] Recall that the TEXTFOOLER texts were selected to be ones that were "particularly bad," i.e., had high predicted score from our regressor, and so we would expect this to be higher than the mean for all TEXTFOOLER texts (2.24) from Table 5.

Table 14
Three scenarios consisting of an original sentence, a TextFooler (TF) variant, and a SusConstrained (SC) variant. For each sentence, the human Likert score ("Human")—not available for SC, by definition—and the regressor-predicted score ("Regr.") are provided. In Scenarios #1 and #2, SusConstrained was chosen over TextFooler, based on votes. In Scenario #3, TextFooler was chosen to be a better variant.

	Scenario #1: SusConstrained > TextFooler: 5 − 0 votes	Human	Regr.
Orig	the skirmishes for power waged among victims and predators settle into an undistinguished rhythm of artificial suspense	1	1.76
TF	the intermingled for power waged among victims and predators settle into an undistinguished rhythm of synthetic awaited	4	3.36
SC	the struggles for power realized with victims and predators settle into an undistinguished rhythm of synthetic suspense		1.83
	Scenario #2: SusConstrained > TextFooler: 3 − 2 votes		
Orig	oliveira seems to pursue silent film representation with every mournful composition.	2	2.13
TF	oliveira seems to sues silent image representation with every mournful composition.	4	2.76
SC	oliveira seems to undertaken silent image representation with every mournful composition.	–	2.46
	Scenario #3: SusConstrained < TextFooler: 2 − 3 votes		
Orig	it's absolutely amazing how first-time director kevin donovan managed to find something new to add to the canon of chan. make chan's action sequences boring.	2	2.51
TF	it's absolutely amazing how first-time director kevin donovan managed to find something new to add to the canon of chan. make chan's action sequences uninspiring.	2	3.17
SC	it's absolutely amazing how first-time director kevin donovan managed to find something new to add to the canon of chan. make chan's action sequences tedium	–	2.38

6. Limitations

A key limitation in this work is that we considered only character-level and word-level attacks, which are available via the *TextAttack* implementation; these types of attacks are well established and have been around for a number of years. Paraphrase-level adversarial attacks represent an intriguing future area of research, especially with the advent of LLMs. However, there are as yet only a very limited number of them, and none of them are suitable for our needs due to various limitations, as discussed in Section 2.1. Further, if they were to have non-suspicion constraints applied, they would either need to be couched within a framework like *TextAttack*, or some other means of application would have to be found. Nevertheless, even though one might expect paraphrase attacks to produce more human-like and consequently less detectable texts, paraphrase work like Lei et al. (2022) demonstrates in their human evaluation that their paraphrases were less grammatical than human-authored sentences, and performed similarly to the word-based attack CLARE. This suggests our notion of graded suspicion could potentially be relevant for paraphrastic attacks in the future as well.

In addition, like the work of Morris et al. (2020a) and Dyrmishi, Ghamizi, and Cordy (2023) that we use as our starting point, this work has only considered sentiment data, and only for English. This is relatively easy to attack; on the other hand, there is perhaps more scope for variation in the quality of adversarial attacks for more challenging tasks, suggesting suspicion measures may be even more relevant on other datasets. Further, we have only explored adversarial attacks within a classification paradigm, rather than generation (Zou et al. 2023), for example.

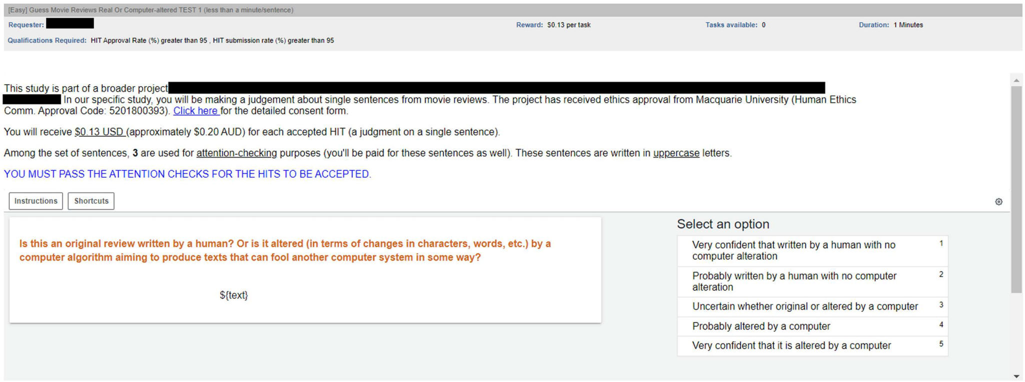
7. Conclusions

In this work, we investigated human judgments of the suspiciousness of adversarial texts, moving beyond binary classifications to capture varying levels of suspicion through a nuanced, Likert-scale approach; as part of this, we have made available a novel dataset of suspiciousness judgments across a range of adversarial methods. Additionally, we examined a range of regression-based models for predicting these judgments, to serve as a baseline for future research on automatically determining how suspicious humans will find adversarial text. We further illustrate the utility of this scorer in imposing suspicion thresholds to enhance the generation of less suspicious adversarial texts.

There are many avenues for future work in understanding and capturing human notions of what makes adversarial text suspicious. As noted above in the Limitations section, extending adversarial methods to LLMs and tasks—both classification tasks like Natural Language Inference used in the adversarial attack literature (Tonni and Dras 2023), or broader tasks like Named Entity Recognition (Srinivasan and Vajjala 2023)—would be important. Beyond this, other kinds of judgments, such as multiple binary judgments gathered on a large scale, could lead to improved models for prediction. More recent LLM work, such as chain-of-thought prompting, could form an alternative approach to scoring, or at least to useful numerical features to use in the sorts of regression models presented in this article. Finally, there is much work to be done to understand how these sorts of scores can be used as constraints to produce more human-like adversarial examples.

Appendix A. Human Evaluation Instruction

The Amazon Mechanical Turk survey interface for human evaluation is illustrated in Figure A.1.



[Easy] Guess Movie Reviews Real Or Computer-altered TEST 1 (less than a minute/sentence)

Requester: [redacted] Reward: $0.13 per task Tasks available: 0 Duration: 1 Minutes

Qualifications Required: HIT Approval Rate (%) greater than 95 , HIT submission rate (%) greater than 95

This study is part of a broader project [redacted] In our specific study, you will be making a judgement about single sentences from movie reviews. The project has received ethics approval from Macquarie University (Human Ethics Comm. Approval Code: 5201800393). Click here for the detailed consent form.

You will receive $0.13 USD (approximately $0.20 AUD) for each accepted HIT (a judgment on a single sentence).

Among the set of sentences, **3** are used for attention-checking purposes (you'll be paid for these sentences as well). These sentences are written in uppercase letters.

YOU MUST PASS THE ATTENTION CHECKS FOR THE HITS TO BE ACCEPTED.

Is this an original review written by a human? Or is it altered (in terms of changes in characters, words, etc.) by a computer algorithm aiming to produce texts that can fool another computer system in some way?

${text}

Select an option

Very confident that written by a human with no computer alteration	1
Probably written by a human with no computer alteration	2
Uncertain whether original or altered by a computer	3
Probably altered by a computer	4
Very confident that it is altered by a computer	5

Figure A.1
Amazon Mechanical Turk UI for collecting human judgments.

Appendix B. ChatGPT Prompt

The prompt used for the ChatGPT baseline is in Table B.1. A few suspicion-scoring responses on the original and adversarial texts are illustrated in Table B.2.

Table B.1
Used prompt for the OpenAI ChatGPT annotations.

"You are a highly intelligent AI trained to detect if a sentence might have been
altered by a computer algorithm."
"In the sentence, one or more words or characters might be altered by a computer
thm to change another algorithm's prediction on it."
"However, sometimes character-level errors can be human errors as well."
"So when judging a sentence you need to see if the sentence raises any suspicion
looking at whether the words go with each other. On a scale of 1–5, where,"
"1 is "no computer alterations",
2 is "possibly no computer alteration",
3 is "uncertain whether altered or not",
4 is "possibly altered" and
5 is "certainly computer-altered",
you will generate a score for the suspiciousness of a sentence."
"For example, "it may . . . wok as a jaunt down memory lane for teens and
young adults" has a score of 4 as there is a character attack on the word "work"
and the letter "r" is deleted, but also can be a human error."
"Another example is "a cliched and surface cautionary tale about the hard-partying
lives of teenagers ." having a score of 5 as the word "shallow" is replaced by another
synonymous word "surface" and "surface cautionary" doesn't make any
sense in this context."
"Another example is "This is a great day to be alive." having a score of 1 as the characters
and words are in place in a meaningful order."
"Produce a numeric score of suspiciousness and reasoning in one line for a given sentence."

Table B.2
LLM (GPT-4) generated suspiciousness scores on two original sentences and two TEXTFOOLER sentences.

Original # a romantic comedy yes but one with characters who think and talk about their goals and are working on hard decisions . **ChatGPT-response #** Score: 1. The sentence appears to be perfectly logical, and all words and characters are used appropriately in the context provided. There is no reason to suspect any alterations were made by an algorithm.
Original # a film that will probably please people already fascinated by behan but leave everyone else yawning with admiration . **ChatGPT-response #** Score: 4. The phrase “yawning with admiration” seems suspicious because it combines two acts that are usually mutually exclusive. One does not generally yawn when they admire something. This could likely be a computer alteration, replacing a word with “yawning”, but it could also be a less common choice of words from a human reviewer
TEXTFOOLER # a disturbing and frighteningly plaintive (orig:evocative) assembly of imagery and melodious (orig:hypnotic) music composed by philip glass . **ChatGPT-response #** Score: 1. The sentence is intact and correctly ordered. There isn’t any out-of-place word, character or synonymous words replacing critical phrases that could raise suspicions of a computer-altered attack.
TEXTFOOLER # the intermingled (orig:skirmishes) for power waged among victims and predator settle into an undistinguished rhythm of synthetic awaited (orig:artificial suspense) . **ChatGPT-response #** I would give this sentence a score of 5. The sentence is full of seemingly unrelated words strung together in a nonsensical sequence, suggesting it was certainly altered by a computer algorithm. Especially the part “synthetic awaited” doesn’t make sense in this context, indicating a high probability of computer alteration.

Acknowledgments

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

Aghababaeyan, Zohreh, Manel Abdellatif, Lionel Briand, Ramesh S., and Mojtaba Bagherzadeh. 2023. Black-box testing of deep neural networks through test case diversity. *IEEE Transactions on Software Engineering*, 49(5):3182–3204. `https://doi.org/10.1109/TSE.2023.3243522`

Alzantot, Moustafa, 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. `https://doi.org/10.18653/v1/D18-1316`

Asl, Javad Rafiei, Mohammad H. Rafiei, Manar Alohaly, and Daniel Takabi. 2024. A semantic, syntactic, and context-aware natural language adversarial example generator. *IEEE Transactions on Dependable and Secure Computing*, pages 1–17. `https://doi.org/10.1109/TDSC.2024.3359817`

Banerjee, Satanjeev and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72.

Bao, Rongzhou, Jiayi Wang, and Hai Zhao. 2021. Defending pre-trained language models from adversarial word substitutions without performance sacrifice. *CoRR*, abs/2105.14553. `https://doi.org/10.18653/v1/2021.findings-acl.287`

Beauchemin, David, Horacio Saggion, and Richard Khoury. 2023. MeaningBERT: Assessing meaning preservation between sentences. *Frontiers in Artificial Intelligence*, 6:1223924. `https://doi.org/10.3389/frai.2023.1223924`, PubMed: 37808622

Belinkov, Yonatan and Yonatan Bisk. 2018. Synthetic and natural noise both break neural machine translation. In *International Conference on Learning Representations*.

Carlini, Nicholas and David A. Wagner. 2017. Towards evaluating the robustness of neural networks. In *2017 IEEE Symposium on Security and Privacy, SP 2017*, pages 39–57. `https://doi.org/10.1109/SP.2017.49`

Cer, Daniel, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017. SemEval-2017 Task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1–14. `https://doi.org/10.18653/v1/S17-2001`

Cer, Daniel, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Brian Strope, and Ray Kurzweil. 2018. Universal sentence encoder for English. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 169–174. `https://doi.org/10.18653/v1/D18-2029`

Chelba, Ciprian, Tomás Mikolov, Mike Schuster, Qi Ge, Thorsten Brants, Phillipp Koehn, and Tony Robinson. 2014. One billion word benchmark for measuring progress in statistical language modeling. In *INTERSPEECH 2014, 15th Annual Conference of the International Speech Communication Association*, pages 2635–2639. `https://doi.org/10.21437/Interspeech.2014-564`

Chen, Yangyi, Hongcheng Gao, Ganqu Cui, Fanchao Qi, Longtao Huang, Zhiyuan Liu, and Maosong Sun. 2022. Why should adversarial perturbations be imperceptible? Rethink the research paradigm in adversarial NLP. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11222–11237. `https://doi.org/10.18653/v1/2022.emnlp-main.771`

Chi, Ryan Andrew, 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. `https://doi.org/10.18653/v1/2022.eval4nlp-1.5`

Chiang, Cheng Han and Hung-yi Lee. 2023. Can large language models be an alternative to human evaluations? In

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 15607–15631. `https://doi.org/10.18653/v1/2023.acl-long.870`

Clark, Elizabeth, Tal August, Sofia Serrano, Nikita Haduong, Suchin Gururangan, and Noah A. Smith. 2021. All that's 'human' is not gold: Evaluating human evaluation of generated text. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7282–7296. `https://doi.org/10.18653/v1/2021.acl-long.565`

Cohen, Gilad, Guillermo Sapiro, and Raja Giryes. 2020. Detecting adversarial samples using influence functions and nearest neighbors. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020*, pages 14441–14450. `https://doi.org/10.1109/CVPR42600.2020.01446`

Dyrmishi, Salijona, 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. `https://doi.org/10.18653/v1/2023.acl-long.491`

Ebrahimi, Javid, 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. `https://doi.org/10.18653/v1/P18-2006`

Eger, Steffen 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. `https://doi.org/10.18653/v1/2020.aacl-main.79`

Gao, Ji, 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*, pages 50–56. `https://doi.org/10.1109/SPW.2018.00016`

Garg, Siddhant 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. `https://doi.org/10.18653/v1/2020.emnlp-main.498`

Gilmer, Justin, Ryan P. Adams, Ian J. Goodfellow, David G. Andersen, and George E. Dahl. 2018. Motivating the rules of the game for adversarial example research. *CoRR*, abs/1807.06732.

Goodfellow, Ian J., Jonathon Shlens, and Christian Szegedy. 2015. Explaining and harnessing adversarial examples. In *3rd International Conference on Learning Representations, ICLR 2015*.

Groh, Matthew, Aruna Sankaranarayanan, Nikhil Singh, Dong Young Kim, Andrew Lippman, and Rosalind Picard. 2024. Human detection of political speech deepfakes across transcripts, audio, and video. *Nature Communications*, 15(1):7629. `https://doi.org/10.1038/s41467-024-51998-z`, PubMed: 39223110

Gu, Ken and Akshay Budhkar. 2021. A package for learning on tabular and text data with transformers. In *Proceedings of the Third Workshop on Multimodal Artificial Intelligence*, pages 69–73. `https://doi.org/10.18653/v1/2021.maiworkshop-1.10`

Han, Xu, Ying Zhang, Wei Wang, Bin Wang, and Yanhui Guo. 2022. Text adversarial attacks and defenses: Issues, taxonomy, and perspectives. *Security and Communication Networks*, 2022:6458488. `https://doi.org/10.1155/2022/6458488`

Huang, Xiaowei, Daniel Kroening, Wenjie Ruan, James Sharp, Youcheng Sun, Emese Thamo, Min Wu, and Xinping Yi. 2020. A survey of safety and trustworthiness of deep neural networks: Verification, testing, adversarial attack and defence, and interpretability. *Computer Science Review*, 37:100270. `https://doi.org/10.1016/j.cosrev.2020.100270`

Ippolito, Daphne, Daniel Duckworth, Chris Callison-Burch, and Douglas Eck. 2020. Automatic detection of generated text is easiest when humans are fooled. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1808–1822. `https://doi.org/10.18653/v1/2020.acl-main.164`

Iyyer, Mohit, John Wieting, Kevin Gimpel, and Luke Zettlemoyer. 2018. Adversarial example generation with syntactically controlled paraphrase networks. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1875–1885. `https://doi.org/10.18653/v1/N18-1170`

Jia, Robin, Aditi Raghunathan, Kerem Göksel, and Percy Liang. 2019. Certified robustness to adversarial word substitutions. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4129–4142. `https://doi.org/10.18653/v1/D19-1423`

Jiang, Haoming, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Tuo Zhao. 2020. SMART: Robust and efficient fine-tuning for pre-trained natural language models through principled regularized optimization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2177–2190. `https://doi.org/10.18653/v1/2020.acl-main.197`

Jin, Di, 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*, pages 8018–8025. `https://doi.org/10.1609/aaai.v34i05.6311`

Kang, Dongyeop, Tushar Khot, Ashish Sabharwal, and Eduard Hovy. 2018. AdvEntuRe: Adversarial training for textual entailment with knowledge-guided examples. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2418–2428. `https://doi.org/10.18653/v1/P18-1225`

Karpinska, Marzena, Nader Akoury, and Mohit Iyyer. 2021. The perils of using Mechanical Turk to evaluate open-ended text generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1265–1285. `https://doi.org/10.18653/v1/2021.emnlp-main.97`

Koh, Pang Wei and Percy Liang. 2017. Understanding black-box predictions via influence functions. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017*, pages 1885–1894.

Kwon, Hyun and Sanghyun Lee. 2022. Ensemble transfer attack targeting text classification systems. *Computer Security*, 117:102695. `https://doi.org/10.1016/J.COSE.2022.102695`

Lei, Qi, Lingfei Wu, Pin-Yu Chen, Alex Dimakis, Inderjit S. Dhillon, and Michael J. Witbrock. 2019. Discrete adversarial attacks and submodular optimization with applications to text classification. In *Proceedings of Machine Learning and Systems*, volume 1, pages 146–165.

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

Levari, David E., Daniel T. Gilbert, Timothy D. Wilson, Beau Sievers, David M. Amodio, and Thalia Wheatley. 2018. Prevalence-induced concept change in human judgment. *Science*, 360(6396):1465–1467. `https://doi.org/10.1126/science.aap8731`, PubMed: 29954981

Li, Dianqi, 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. `https://doi.org/10.18653/v1/2021.naacl-main.400`

Li, Guoyi, 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. `https://doi.org/10.18653/v1/2023.findings-emnlp.1053`

Li, Jinfeng, 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*. `https://doi.org/10.14722/ndss.2019.23138`

Li, Linyang, 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. `https://doi.org/10.18653/v1/2020.emnlp-main.500`

Lin, Chin Yew. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81.

Liu, Han, 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.

Liu, Na, Mark Dras, and Wei Emma Zhang. 2022. Detecting textual adversarial examples based on distributional characteristics of data representations. In *Proceedings of the 7th Workshop on Representation Learning for NLP*, pages 78–90. `https://doi.org/10.18653/v1/2022.repl4nlp-1.9`

Ma, Xingjun, 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*.

Mai, Kimberly T., Sergi Bray, Toby Davies, and Lewis D. Griffin. 2023. Warning: Humans cannot reliably detect speech deepfakes. *PLoS One*, 218(8):e0285333. `https://doi.org/10.1371/journal.pone.0285333`, PubMed: 37531336

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc.

Moosavi-Dezfooli, Seyed-Mohsen, Alhussein Fawzi, and Pascal Frossard. 2016. DeepFool: A simple and accurate method to fool deep neural networks. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016*, pages 2574–2582. `https://doi.org/10.1109/CVPR.2016.282`

Morris, John, 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. `https://doi.org/10.18653/v1/2020.findings-emnlp.341`

Morris, John, 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. `https://doi.org/10.18653/v1/2020.emnlp-demos.16`

Mosca, Edoardo, Shreyash Agarwal, Javier Rando Ramírez, and Georg Groh. 2022. “That is a suspicious reaction!”: Interpreting logits variation to detect NLP adversarial attacks. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7806–7816. `https://doi.org/10.18653/v1/2022.acl-long.538`

Mrkšić, Nikola, Diarmuid Ó Séaghdha, Blaise Thomson, Milica Gašić, Lina M. Rojas-Barahona, Pei-Hao Su, David Vandyke, Tsung-Hsien Wen, and Steve Young. 2016. Counter-fitting word vectors to linguistic constraints. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 142–148. `https://doi.org/10.18653/v1/N16-1018`

Nguyen-Son, Hoang Quoc, Seira Hidano, Kazuhide Fukushima, Shinsaku Kiyomoto, and Isao Echizen. 2023. VoteTRANS: Detecting adversarial text without training by voting on hard labels of transformations. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5090–5104. `https://doi.org/10.18653/v1/2023.findings-acl.315`

O’Mahony, Michael. 1986. The binomial test: Applications in sensory difference and preference testing. In *Sensory Evaluation of Food*, pages 57–88. `https://doi.org/10.1201/9780203739884-5`

Pang, Bo 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. `https://doi.org/10.3115/1219840.1219855`

Papernot, Nicolas and Patrick D. McDaniel. 2018. Deep k-nearest neighbors: Towards confident, interpretable and robust deep learning. *CoRR*, abs/1803.04765.

Papernot, Nicolas, Patrick D. McDaniel, Somesh Jha, Matt Fredrikson, Z. Berkay Celik, and Ananthram Swami. 2016. The limitations of deep learning in adversarial settings. In *IEEE European Symposium on Security and Privacy, EuroS&P 2016*, pages 372–387. `https://doi.org/10.1109/EuroSP.2016.36`

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318. `https://doi.org/10.3115/1073083.1073135`

Peng, Hao, Shixin Guo, Dandan Zhao, Xuhong Zhang, Jianmin Han, Shouling Ji, Xing Yang, and Ming Zhong. 2023. TextCheater: A query-efficient textual adversarial attack in the hard-label setting. *IEEE Transactions on Dependable and Secure Computing*, pages 1–16. `https://doi.org/10.1109/TDSC.2023.3339802`

Pennington, Jeffrey, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543. `https://doi.org/10.3115/v1/D14-1162`

Pruthi, Danish, 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. `https://doi.org/10.18653/v1/P19-1561`

Qi, Fanchao, Yangyi Chen, Xurui Zhang, Mukai Li, Zhiyuan Liu, and Maosong Sun. 2021. Mind the style of text! Adversarial and backdoor attacks based on text style transfer. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4569–4580. `https://doi.org/10.18653/v1/2021.emnlp-main.374`

Ren, Shuhuai, 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. `https://doi.org/10.18653/v1/P19-1103`

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

Ribeiro, Marco Tulio, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912. `https://doi.org/10.18653/v1/2020.acl-main.442`

Sellam, Thibault, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892. `https://doi.org/10.18653/v1/2020.acl-main.704`

Srinivasan, Akshay and Sowmya Vajjala. 2023. A multilingual evaluation of NER robustness to adversarial inputs. In *Proceedings of the 8th Workshop on Representation Learning for NLP (RepL4NLP 2023)*, pages 40–53. `https://doi.org/10.18653/v1/2023.repl4nlp-1.4`

Sundar, S. Shyam, Maria D. Molina, and Eugene Cho. 2021. Seeing is believing: Is video modality more powerful in spreading fake news via online messaging apps? *Journal of Computer-Mediated Communication*, 26(6):301–319. `https://doi.org/10.1093/jcmc/zmab010`

Szegedy, Christian, 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*.

Tian, Junfeng, Zhiheng Zhou, Man Lan, and Yuanbin Wu. 2017. ECNU at SemEval-2017 Task 1: Leverage kernel-based traditional NLP features and neural networks to build a universal model for multilingual and cross-lingual semantic textual similarity. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 191–197. `https://doi.org/10.18653/v1/S17-2028`

Tonni, Shakila Mahjabin and Mark Dras. 2023. What learned representations and influence functions can tell us about adversarial examples. In *Findings of the Association for Computational Linguistics: IJCNLP-AACL 2023 (Findings)*, pages 392–411. https://doi.org/10.18653/v1/2023.findings-ijcnlp.35

van den Heuvel, Edwin and Zhuozhao Zhan. 2022. Myths about linear and monotonic associations: Pearson's r, Spearman's ρ, and Kendall's τ. *The American Statistician*, 76(1):44–52. https://doi.org/10.1080/00031305.2021.2004922

Vogel, Carl, 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*. Springer, pages 383–393. https://doi.org/10.1007/978-981-13-8950-4_34

Wang, Alex, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355. https://doi.org/10.18653/v1/W18-5446

Wang, Boxin, Chejian Xu, Shuohang Wang, Zhe Gan, Yu Cheng, Jianfeng Gao, Ahmed Hassan Awadallah, and Bo Li. 2021a. Adversarial GLUE: A multi-task benchmark for robustness evaluation of language models. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021*.

Wang, Jiongxiao, Zichen Liu, Keun Hee Park, Muhao Chen, and Chaowei Xiao. 2023a. Adversarial demonstration attacks on large language models. *CoRR*, abs/2305.14950. https://doi.org/10.48550/ARXIV.2305.14950

Wang, Lihao and Xiaoqing Zheng. 2020. Improving grammatical error correction models with purpose-built adversarial examples. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2858–2869. https://doi.org/10.18653/v1/2020.emnlp-main.228

Wang, Tianlu, 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. https://doi.org/10.18653/v1/2020.emnlp-main.417

Wang, Xiaosen, Jin Hao, Yichen Yang, and Kun He. 2021b. Natural language adversarial defense through synonym encoding. In *Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence, UAI 2021*, volume 161 of *Proceedings of Machine Learning Research*, pages 823–833.

Wang, Xiaosen, Yifeng Xiong, and Kun He. 2022. Detecting textual adversarial examples through randomized substitution and vote. In *Uncertainty in Artificial Intelligence, Proceedings of the Thirty-Eighth Conference on Uncertainty in Artificial Intelligence*, volume 180 of *Proceedings of Machine Learning Research*, pages 2056–2065.

Wang, Zimu, Wei Wang, Qi Chen, Qiufeng Wang, and Anh Nguyen. 2023b. Generating valid and natural adversarial examples with large language models. *CoRR*, abs/2311.11861.

Xiao, Mingxuan, Yan Xiao, Hai Dong, Shunhui Ji, and Pengcheng Zhang 2023. LEAP: Efficient and automated test method for NLP software. In *38th IEEE/ACM International Conference on Automated Software Engineering, ASE 2023*, pages 1136–1148. https://doi.org/10.1109/ASE56229.2023.00052

Yildiz, Eray 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. https://doi.org/10.18653/v1/W19-4205

Yoo, Jin Yong and Yanjun Qi. 2021. Towards improving adversarial training of NLP models. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 945–956. https://doi.org/10.18653/v1/2021.findings-emnlp.81

You, Wencong, Zayd Hammoudeh, and Daniel Lowd. 2023. Large language models are better adversaries: Exploring generative clean-label backdoor attacks against text classifiers. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12499–12527. https://doi.org/10.18653/v1/2023.findings-emnlp.833

Zang, Yuan, Fanchao Qi, Chenghao Yang, Zhiyuan Liu, Meng Zhang, Qun Liu, and

Maosong Sun. 2020. Word-level textual adversarial attacking as combinatorial optimization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6066–6080. `https://doi.org/10.18653/v1/2020.acl-main.540`

Zhang, Jinghao, Yuting Liu, Qiang Liu, Shu Wu, Guibing Guo, and Liang Wang. 2024. Stealthy attack on large language model based recommendation. *CoRR*, abs/2402.14836. `https://doi.org/10.48550/ARXIV.2402.14836`

Zhang, Tianyi, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020*.

Zhang, Wei Emma, Quan Z. Sheng, Ahoud Alhazmi, and Chenliang Li. 2020. Adversarial attacks on deep-learning models in natural language processing: A survey. *ACM Transactions on Intelligent Systems and Technology*, 11(3):1–41. `https://doi.org/10.1145/3374217`

Zheng, Boyuan, Xiaoyu Yang, Yu-Ping Ruan, Zhenhua Ling, Quan Liu, Si Wei, and Xiaodan Zhu. 2021. SemEval-2021 Task 4: Reading comprehension of abstract meaning. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 37–50. `https://doi.org/10.18653/v1/2021.semeval-1.4`

Zhu, Hai, Qingyang Zhao, Weiwei Shang, Yuren Wu, and Kai Liu. 2024. LimeAttack: Local explainable method for textual hard-label adversarial attack. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014*, pages 19759–19767. `https://doi.org/10.1609/aaai.v38i17.29950`

Zou, Andy, Zifan Wang, Nicholas Carlini, Milad Nasr, J. Zico Kolter, and Matt Fredrikson. 2023. Universal and transferable adversarial attacks on aligned language models. *ArXiv preprint arXiv:2307.15043*.