Adversarial Text Purification: A Large Language Model Approach for Defense

Raha Moraffah, Shubh Khandelwal, Amrita Bhattacharjee, and Huan Liu

Arizona State University, Tempe, AZ, USA
{rmoraffa,skhand15,abhatt43,huanliu}@asu.edu

Abstract. Adversarial purification is a defense mechanism for safeguarding classifiers against adversarial attacks without knowing the type of attacks or training of the classifier. These techniques characterize and eliminate adversarial perturbations from the attacked inputs, aiming to restore purified samples that retain similarity to the initially attacked ones and are correctly classified by the classifier. Due to the inherent challenges associated with characterizing noise perturbations for discrete inputs, adversarial text purification has been relatively unexplored. In this paper, we investigate the effectiveness of adversarial purification methods in defending text classifiers. We propose a novel adversarial text purification that harnesses the generative capabilities of Large Language Models (LLMs) to purify adversarial text without the need to explicitly characterize the discrete noise perturbations. We utilize prompt engineering to exploit LLMs for recovering the purified samples for given adversarial examples such that they are semantically similar and correctly classified. Our proposed method demonstrates remarkable performance over various classifiers, improving their accuracy under the attack by over 65% on average.

Keywords: Textual Adversarial Defenses · Adversarial Purification · Textual Adversarial Defenses· Large Language Model.

1 Introduction

Despite the tremendous success of text classification models [9][20], studies have exposed their susceptibility to adversarial examples, i.e., carefully crafted sentences with human-unrecognizable changes to the inputs that are misclassified by the classifiers [13]. The dependability and integrity of NLP applications are seriously threatened by the vulnerability of text classification models to these attacks. Thus, developing stronger defenses against adversarial attacks is crucial in improving the classification model's robustness.

Adversarial purification is a type of defense mechanism against adversarial attacks. It characterizes and removes the adversarial perturbations from the attacked inputs to generate purified samples that are similar to the attacked ones and are classified correctly by the classifier [25][30][36][31]. These methods have demonstrated efficacy in the field of image classification without making

assumptions on the form of an attack and a classification model, thus being able to defend pre-existing classifiers against unseen threats. The potential of adversarial purification, however, has not been explored for text classification, due to the challenges of characterizing the adversarial perturbations for discrete data. In particular, contrary to images, where perturbations can be generated based on continuous gradients, for text data, adversarial perturbations are generated by manipulating combinations of words in the input text [13]. Therefore, identifying these perturbations is also a combinatorial problem.

An ideal solution to adversarial purification for text is to generate the purified example without explicitly characterizing the noise perturbations. In an attempt to achieve this, Li et al. [18] propose a greedy approach that randomly masks the adversarial examples and uses their reconstructed versions by the Masked Language Models (e.g., BERT [9]) as benign purified examples. However, due to its greedy nature, this defense can be ineffective for defending text classifiers.

The exponential growth of the sheer size of LLMs has expedited their generative applications in various fields [27]. To study the effectiveness of adversarial purification for texts, we investigate if LLMs can be exploited to directly generate the purified examples from their adversarial counterparts, eliminating the need for the characterization of adversarial perturbation. To this end, we utilize the generative power of instruction-based LLMs, particularly GPT-3.5, and design a prompt to exploit the contextual understanding and capacity of LLMs to recover purified samples.

Compared to the greedy approach of selecting random combinations of tokens iteratively to remove adversarial perturbations, our proposed method exploits the comprehension and contextual understanding of LLMs to effectively reverse the adversarial perturbations, while utilizing their extensive generation power and capacity to produce cohesive, fluent texts. Our method demonstrates the effective use of adversarial purification methods for text classification, improving the performance of the classifier under attack by over 65%, and improving the performance of the existing text purification defense by over 25% in most cases. Our results open a new avenue for future research in textual adversarial defense based on purification. Our contributions are summarized as follows:

- We study if it is possible to effectively implement the adversarial text purification defense for text.
- We are the first to utilize the contextual understanding and capacity of LLMs for effective text-based adversarial purification defense.
- We conduct extensive experiments on two state-of-the-art transformer-based text classifiers and demonstrate the effectiveness of our proposed adversarial purification method in defending the pre-trained classifiers against strong attacks without any knowledge of the attack.

2 Related Work

Adversarial attacks on text classifiers: Over the years there have been various types of adversarial attacks for text, with varying degrees of success on different types of model architectures. Adversarial attacks, broadly categorized into black box and white box [32], manipulate textual data through insertion, deletion, or swapping of characters and words. The substitution-based strategies to craft adversarial examples employ techniques like genetic algorithms, greedy-search, or gradient-based methods for word replacement [2, 13, 29]. Recent works involving word-level perturbations include TextFooler [13], BERT-Attack [16], TextHoaxer [35]. Alongside the vast bidy of work on word-level attacks, there is also significant amount of works in character-level and sentence-level attacks [32].

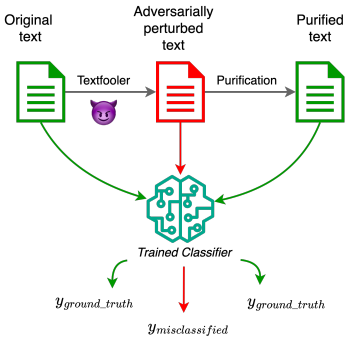

Fig. 1. Our Proposed LLM-guided Adversarial Text Purification Framework.

Adversarial purification & other defenses: Influenced by the rapid development of various adversarial attacks in text, there has also been an increasing number of defense mechanisms to ensure robustness of models against different types of attacks. Some of these defense methods introduce certified robust models to create a defensive range within which substitutions cannot perturb the model [12]. Gradient-based adversarial training strategies have shown effectiveness in defending attacks with no prior knowledge and improving defense [23, 22, 10, 8, 38, 17]. Adversarial purification is a particularly desirable type of defense since it does not require prior knowledge of the type of attack. Prior work in adversarial purification has traditionally focused on continuous inputs [18] such as images, exploring generative models such as GANs [30], EBMs [15], and diffusion models [33, 25]. However, the field of creating better adversarial defenses and improving robustness in NLP has experienced considerable interest in recent years. Adversarial purification has been explored, however, it is comparably uncommon in NLP. [18] aims to utilize the contextual and masking capabilities of pre-trained masked language models (such as BERT [9]) in order to create a defense against adversarial attacks. However, in this work, we aim to use the power of generative AI, in particular, recent state-of-the-art Large Language Models (LLMs) to perform adversarial purification in the context of capabilities to explore the possibility of improving the robustness of the models.

LLMs as pseudo-oracles: Alongside the impressive performance of LLMs on a variety of natural language tasks [7], LLMs are also being increasingly used as pseudo-oracles, such as in data annotation [11, 1, 14], as detectors [4], for model explainability [5] and as experts in general [34]. Inspired by such works, in this work, we propose to use LLMs to perform adversarial purification in the challenging text domain.

3 Background

3.1 Large Language Models

Large language models (LLMs) are essentially deep networks that are based on transformer networks. Transformer-based LLMs are highly effective models that are capable of learning and generating natural language. Broadly there are two categories of language models: (i) Autoregressive language models and (ii) Masked language models. Autoregressive language models are simply trained to predict the next token in a sentence, thereby learning how to generate fluent text when pre-trained on a large corpora of data. Such models include GPT-2 [28], GPT-3 [6], etc. Masked language models (MLMs) are bi-directional models that learn by first masking some fraction of tokens in the sentence and then predicting appropriate tokens to fill the masked slots. Examples of such models include BERT [9], RoBERTa [20], etc. The bidirectional nature of MLMs help the models to have higher language understanding capabilities, and thereby better performance on NLU tasks. More recently, autoregressive models such as the GPT-3 family of models are also being further trained via instruction-tuning [26] with (instruction, response) text pairs, whereby the model learns to generate text to follow user-specified instructions and perform tasks. Some of these instruction-tuned models undergo further training steps (e.g., via RLHF [3]) to align their responses with human preferences. State-of-the-art LLMs such as GPT-3.5 and GPT-4 from OpenAI demonstrate impressive performance when it comes to understanding long and complex human-written instructions in the prompts, as well as editing and generating text. Therefore, we use one of these models in an off-the-shelf manner for our framework.

3.2 Adversarial Text Purification

Adversarial purification is an adversarial defense mechanism that is relatively newer in the natural language domain. As we elaborated in the previous section, this method has been well explored in the domain of computer vision, whereby generative models are used to perform the purification. In the image domain, the standard method is to inject random noise into a perturbed input image, and then use a generative model i.e., the purification algorithm to reconstruct the original *clean* image from the noisy image over multiple rounds. The generated image would now be free of the adversarial perturbations. However, in the domain of text, the discrete nature of the input makes it infeasible to apply the standard computer vision methods directly. One recent attempt at adversarial text purification [18] uses masked language models to randomly mask multiple copies of the perturbed text, and then recovering the text by filling in the mask using the masked language model. This method essentially is somewhat similar to the standard process of injecting noise and iteratively reconstructing the input, as followed in the image domain. However, there is no other method for performing adversarial text purification. To fill this gap, we propose to directly leverage the instruction understanding and text generation capabilities of recent state-of-the-art LLMs and use these LLMs to perform the text purification.

4 LLM-guided Adversarial Text Purification

In this section, we describe our purification framework and explain the necessary design choices.

We show our overall framework in Figure 1. As mentioned previously, in this work we focus on the task of text classification and we use fine-tuned pre-trained language models (such as BERT [9]), denoted by $f(\cdot)$ as the classifier. During inference, we evaluate such a classifier on the test set of our task dataset (X_{test}, Y_{test}) where X_{test} and Y_{test} are the sequence of input texts and associated ground truth labels respectively. For an input text $x_i \in X_{test}$, say the classifier correctly predicted $f(x_i) = y_i$, or y_{ground_truth} for ease of reference. Now, say this text is perturbed by an adversarial attack method such that the perturbed text x_i' now gets misclassified to a different label, say $y_{misclassified}$. While many defense mechanisms train the model i.e., the classifier to be adversarially robust to some specific categories of perturbations, purification methods enable simply editing the text, ideally removing the adversarial perturbation from the text and thereby enabling the model to correctly classify the text. Following this, we collect this set of adversarially perturbed input texts X_{test}' and attempt to purify them by using off-the-shelf large language models. In order to do this, we carefully design prompts, as elaborated in the following paragraph. After the purification step, we obtain $\tilde{X}_{test}$ which then is correctly classified by the classifier in majority of the cases.

We use an instruction-tuned LLM which is capable of following human-written instructions in the prompt, in order to generate the purified samples. To enable this, we carefully design the following prompt:

> 'Human: You are a teacher tasked with grading a quiz. The quiz consists of a sentence (the question) and a classification label (the student's answer). Unfortunately, the sentence has been manipulated by an adversarial attack, leading to a misclassification.
> Given the altered sentence and its incorrect label, your job is to generate a new sentence that is semantically similar to the altered one but will be classified correctly according to the correct label.
> The categories for classification are: [list of classification categories]
> ALTERED SENTENCE (QUESTION): [altered sentence]
> MISCLASSIFIED LABEL (STUDENT ANSWER): [misclassified label]
> CORRECT LABEL (TRUE ANSWER): [correct label]
> Please create a new sentence that conveys the same meaning as the altered sentence but will be classified under the CORRECT LABEL when graded .
> Even if there is not a misclassification, provide/construct the sentence to the best of your capability. The output format must be json:
> "Original Sentence": "[New sentence here]" Begin!'

In the prompt above, [altered sentence] refers to the adversarially perturbed input text x_i', [misclassified label] refers to $y_{misclassified}$, [correct label] refers to y_{ground_truth} and [list of classification categories] refer to the list of possible labels for the particular classification task. As evident in the prompt, we 'prime' the LLM to enable it to act like a knowledgable teacher, thereby guiding the editing process. This is the prompt we use for eliciting the purified version of the text from the LLM, and we denote this prompt as `P0`.

To investigate the efficacy of this carefully designed prompt, we further design and test out two variants of this prompt: `P1`: which removes the instruction regarding generating text that would correct the misclassified label, and `P2`: which essentially prompts the LLM to generate a paraphrased version of the input text. The prompt `P1` is created by simply removing the text highlighted in pink from `P0`. Finally, the prompt `P2` is:

> 'Human: Please generate a paraphrased sentence version of the following sentence.
> SENTENCE: [altered sentence]
> The output format must be json:
> "Original Sentence": "[Paraphrased sentence here]" Begin!'

5 Experiments

We conduct comprehensive experiments to evaluate the effectiveness of our proposed LLM-guided adversarial purification method. Our experiments are designed to examine the three main aspects of our method: **(i)** Effectiveness of the proposed method;**(ii)** Ablation study of the components of the designed prompt; and, **(iii)** case study of the purified examples. In the following, we first explain our experimental setting and then discuss our experimental results.

5.1 Experimental Setting

In this section, we describe the datasets, the adversarial attack and the LLM we used in our experiments. We also describe the relevant defense baselines we compare our method to and provide information on our experimental setup to ensure reproducibility. Note that our experimental settings closely follow the ones in the state-of-the-art methods [18].

Datasets. We conduct experiments on two commonly-used benchmark NLP datasets: (1) **IMDb** [21]: for sentiment classification of movie reviews where each review is labeled with a *positive* or *negative* label, and (2) **AG News** [37]: news topic classification where each article is labeled with one of the four categories of {*science, business, world, sports*}.

Adversarial Attack and Defense Baselines. For all our experiments we use the one of the strongest textual attacks named TextFooler [13]. Similar to our baselines, we use the open-source implementation of TextAttack library [24]. The TextFooler attack is selected due to its efficient generation of strong and

Defense ↓	Original Accuracy	TextFooler ($K = 12$)	TextFooler ($K = 50$)
IMDb ↓			
fine-tuned BERT	94.1	20.4	2.8
Adv-HotFlip (BERT)	95.1	36.1	8.0
FreeLB (BERT)	96.0	30.2	7.3
FreeLB++ (BERT)	93.2	-	45.3
Text purification (BERT) [18]	93.0	<u>81.5</u>	51.0
Text purification (RoBERTa) [18]	96.1	**84.2**	54.3
(*Ours*) LLM-guided purification (BERT)	94.54	79.34	<u>73.52</u>
(*Ours*) LLM-guided purification (RoBERTa)	95.06	78.9	**76.16**
AG News ↓			
fine-tuned BERT	92.0	32.8	19.4
Adv-HotFlip (BERT)	91.2	35.3	18.2
FreeLB (BERT)	90.5	40.1	20.1
Text purification (BERT) [18]	90.6	61.5	34.9
Text purification (RoBERTa) [18]	90.8	59.1	34.2
(*Ours*) LLM-guided purification (BERT)	95.12	**83.58**	<u>81.3</u>
(*Ours*) LLM-guided purification (RoBERTa)	94.76	<u>82.84</u>	**81.4**

Table 1. Comparison of our LLM-guided purification methods with baselines as described in Section 5.1. Post-attack accuracy numbers are as reported in [18]. **Bold** denotes the best performance in terms of recovered accuracy, and <u>underline</u> implies the second-best performance.

highly successful adversarial examples, making it an ideal attack to assess the effectiveness of the defense mechanisms. Following previous work, for the size of candidate list we choose $K = \{12, 50\}$ in our experiments.

Following the previous work on adversarial purification [18], we compare the performance of our method with two types of adversarial defense, namely (1) *Textual adversarial training methods:* these methods are based on adversarial training of the classifiers using the adversarial examples generated based on the gradients of the latent space. We use **Adv-HotFlip [10]** and **FreeLB [17]**, two state-of-the-arts in this category that do not require For the choice of baseline defenses, as well as the **FreeLB++ [19]**, which requires the candidate list; and (2) *Textual adversarial purification methods:* methods based on purifying the adversarial examples to generate correctly-classified benign examples. To the best of our knowledge, only one text adversarial purification method exists as in [18]. We include this method as our baseline.

Classifier. Following work in [18] we use classifiers based on two pre-trained masked language models: BERT [9] and RoBERTa [20]. For each dataset, we use BERT and RoBERTa models from Huggingface Transformers (bert-base-

uncased[1] and roberta-base[2]), fine-tuned on that specific dataset. Note that our proposed method does not require any further fine-tuning or adversarial training of the model and we can simple query the fine-tuned BERT and RoBERTa models in an off-the-shelf manner. For evaluating our framework, we report the post-attack accuracy, with and without the purification method, along with the original classifier accuracy without any attack.

Implementation details. We use OpenAI's GPT-3.5 (version as of November 2023) and use our carefully designed prompts to obtain purified versions of adversarially altered texts. The process involved crafting prompts that guide the model to generate semantically similar but unperturbed versions of the input texts. We chose GPT-3.5 for its advanced contextual understanding and generative capabilities as indicated in [6]. We automated this process using the OpenAI API[3] and LangChain[4]. Our experiments were implemented in Pytorch and were run on two systems: (1) Linux system with one A30 and (ii) Linux system with four A100s. All code and links to data will be made available.

Effectiveness of Proposed Purification Method

5.2 Results & Discussion

In this section, we aim to answer if our proposed LLM-based adversarial text purification method is able to effectively purify the adversarial examples. For the sake of comparison, we also report the accuracy under attack for vanilla fine-tuned classifiers. We apply our defense and the state-of-the-art adversarial defenses on the IMDB and AG News datasets and report the results in Table 1. Our results demonstrate that our proposed method effectively defends the state-of-the-art transform-based text classifiers, improving their accuracy under attack by more than 60% in most cases. We elaborate on our observations in the following: (1) The adversarial training-based defenses, i.e., Adv-HotFlip, FreeLB, and FreeLB++, are constantly outperformed by our method based on purification by a large margin (more than 30%). This is because these models are robustified against continuous gradient-based adversarial perturbations and not the discrete word-level perturbations used by text adversarial attacks; (2) the state-of-the-art purification-based defense, namely Text purification, has remarkably lower performance compared to our method. This is because the

Prompt Type	AG News
Original (BERT)	95.12
Full prompt `P0`	**81.3**
`P1`	78
`P2`	52.7

Table 2. Effectiveness of our full prompt as described in Section 4 (denoted by `P0`).

[1] https://huggingface.co/bert-base-uncased
[2] https://huggingface.co/roberta-base
[3] https://platform.openai.com/docs/api-reference
[4] https://www.langchain.com/

Text purification method is based on a greedy approach and iteratively selects and perturbs random words. Our method, on the other hand, utilizes the power of LLMs to directly generate purified examples; and (3) finally, our proposed method (LLM-guided purification) achieves the highest after attack accuracy, which is comparable to the accuracy of the model before the attack. For instance, for the BERT trained on the AG News dataset, the original accuracy before the attack is 95.06%, whereas the accuracy after the attack is 83.58%, which is more that 20% better than the accuracy under attack for the second best-performing defense (Text purification (BERT)).

	Texts	**Label**
Original	E-mail scam targets police chief Wiltshire Police warns about "phishing" after its fraud squad chief was targeted.	science
Adv. Perturbed ($K = 12$)	E-mail scam targets gendarmerie chief Wiltshire Police warns about "phishing" after its deception battalion massa was targeted.	the world
LLM-purified	Wiltshire Police issues warning about phishing email scam targeting their deception battalion massa.	science (conf.: 0.994)
Adv. Perturbed ($K = 50$)	E-mail scam targets police chief Wiltshire Police warns about "phishing" after its hoax battalion leiter was targeted.	the world
LLM-purified	Wiltshire Police alerts about a scam email targeting their police chief, warning about phishing after their hoax battalion leiter was targeted.	science (conf.: 0.984)
Original	Consumer Prices Down, Industry Output Up WASHINGTON (Reuters) - U.S. consumer prices dropped in July for the first time in eight months as a sharp run up in energy costs reversed, the government said in a report that suggested a slow rate of interest rate hikes is likely.	business
Adv. Perturbed ($K = 12$)	Eaters Pricing Down, Departments Product Arriba WASHINGTON (Reuters) - U.S. consuming prices declined in July for the first time in eight months as a ferocious manage up in energy costs quashed, the government tell in a notification that recommendations a sluggish cadence of relevance pace hiking is possible.	science
LLM-purified	Consumers face lower prices as government report suggests slower pace of interest rate hikes due to decrease in energy costs.	business (conf.: 0.954)
Adv. Perturbed ($K = 50$)	User Charging Down, Industry Product Up WASHINGTON (Reuters) - U.S. clients prices dwindled in July for the first time in eight months as a sharp run up in energy costs quashed, the government tell in a report that recommendation a slow rate of interest rate hikes is likely.	science
LLM-purified	U.S. consumer prices fell in July for the first time in eight months due to a significant increase in energy costs, as reported by the government. This suggests that the pace of interest rate hikes is likely to slow down.	business (conf.: 0.999)

Table 3. Examples from the AG News dataset with TextFooler perturbations (with both $K = 12$ and $K = 50$) along with LLM-purified versions of the perturbed input. Portions of the input text altered by the TextFooler method are shown in teal. Labels in blue are correctly classified, labels in red are misclassified. We see that our methods successfully retains the original label after attack, while maintaining semantics of the original input.

Ablation: Effectiveness of Prompt Components In this section, we conduct experiments with two additional prompts namely `P1` and `P2` as explained in Section 4, and compare their results with the results obtained using the main prompt (`P0`). Specifically, `P1` is designed to understand the effect of the explicit instruction to ensure the purified text is classified as the correct label. The goal of designing `P2` is to assess the effectiveness of our proposed prompt to ensure the purified samples retain semantic similarity to the adversarial counterparts. To this end `P2` simply asks the LLM to paraphrase the adversarial example. Our results reported in Table 2 indicate the effectiveness of our main prompt. The accuracy under attack for purification based on `P1` is about 4% less than the full prompt `P0`. This indicates that even though the full prompt is useful to achieve higher performance, our proposed methodology can obtain similar performance, even when the original correct label of the sample is unknown. However, the performance achieved with `P2`is remarkably lower compared to the main prompt, indicating that our proposed prompt is indeed necessary for a successful adversarial purification.

Case study We showcase some examples from the AG News dataset in Table 3. We can observe that our purified examples are semantically similar to the adversarial examples while being classified to the original correct class before the attack. This shows that our method can successfully remove the adversarial perturbation and does not change the original benign content of the example. It is important to note that our method can effectively remove adversarial perturbations of any length with only one prompt. Additionally, our generated examples are fluent and grammatically correct, due to the generative power of the LLMs.

6 Conclusion

In this paper, we propose a novel text adversarial purification method, that can effectively remove the adversarial perturbations of any lengths from the adversarial examples and generate purified examples that are semantically similar but are classified to the original correct class. Overcoming the challenges of characterizing adversarial perturbations for discrete inputs (i.e., text), our proposed method utilizes the advanced contextual understanding and generative capabilities of the LLMs to effectively purify the adversarial examples. More concretely, we employ prompt engineering to leverage Large Language Models (LLMs) in the retrieval of purified examples from provided adversarial instances, ensuring both semantic similarity and accurate classification. Our novel method exhibits impressive performance across diverse classifiers, resulting in an average accuracy improvement of over 65% under adversarial attacks.

7 Acknowledgements

This work is supported by Army Research Office (ARO) W911NF2110030 and Army Research Laboratory (ARL) W911NF2020124. Opinions, interpretations,

conclusions, and recommendations are those of the authors' and should not be interpreted as representing the official views or policies of the Army Research Office or the Army Research Lab.

References

1. Alizadeh, M., Kubli, M., Samei, Z., Dehghani, S., Bermeo, J.D., Korobeynikova, M., Gilardi, F.: Open-source large language models outperform crowd workers and approach chatgpt in text-annotation tasks. arXiv preprint arXiv:2307.02179 (2023)
2. Alzantot, M., Sharma, Y., Elgohary, A., Ho, B.J., Srivastava, M., Chang, K.W.: Generating natural language adversarial examples. arXiv preprint arXiv:1804.07998 (2018)
3. Bai, Y., Jones, A., Ndousse, K., Askell, A., Chen, A., DasSarma, N., Drain, D., Fort, S., Ganguli, D., Henighan, T., et al.: Training a helpful and harmless assistant with reinforcement learning from human feedback. arXiv preprint arXiv:2204.05862 (2022)
4. Bhattacharjee, A., Liu, H.: Fighting fire with fire: Can chatgpt detect ai-generated text? arXiv preprint arXiv:2308.01284 (2023)
5. Bhattacharjee, A., Moraffah, R., Garland, J., Liu, H.: Llms as counterfactual explanation modules: Can chatgpt explain black-box text classifiers? arXiv preprint arXiv:2309.13340 (2023)
6. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. Advances in neural information processing systems **33**, 1877–1901 (2020)
7. Chang, Y., Wang, X., Wang, J., Wu, Y., Zhu, K., Chen, H., Yang, L., Yi, X., Wang, C., Wang, Y., et al.: A survey on evaluation of large language models. arXiv preprint arXiv:2307.03109 (2023)
8. Cheng, Y., Jiang, L., Macherey, W.: Robust neural machine translation with doubly adversarial inputs. arXiv preprint arXiv:1906.02443 (2019)
9. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
10. Ebrahimi, J., Rao, A., Lowd, D., Dou, D.: Hotflip: White-box adversarial examples for text classification. arXiv preprint arXiv:1712.06751 (2017)
11. Flamholz, Z.N., Biller, S.J., Kelly, L.: Large language models improve annotation of viral proteins. Research Square (2023)
12. Jia, R., Raghunathan, A., Göksel, K., Liang, P.: Certified robustness to adversarial word substitutions. arXiv preprint arXiv:1909.00986 (2019)
13. Jin, D., Jin, Z., Zhou, J.T., Szolovits, P.: Is bert really robust? a strong baseline for natural language attack on text classification and entailment. In: Proceedings of the AAAI conference on artificial intelligence. vol. 34, pp. 8018–8025 (2020)
14. Latif, S., Usama, M., Malik, M.I., Schuller, B.W.: Can large language models aid in annotating speech emotional data? uncovering new frontiers. arXiv preprint arXiv:2307.06090 (2023)
15. LeCun, Y., Chopra, S., Hadsell, R., Ranzato, M., Huang, F.: A tutorial on energy-based learning. Predicting structured data **1**(0) (2006)
16. Li, L., Ma, R., Guo, Q., Xue, X., Qiu, X.: Bert-attack: Adversarial attack against bert using bert. arXiv preprint arXiv:2004.09984 (2020)

17. Li, L., Qiu, X.: Token-aware virtual adversarial training in natural language understanding. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 35, pp. 8410–8418 (2021)
18. Li, L., Song, D., Qiu, X.: Text adversarial purification as defense against adversarial attacks. arXiv preprint arXiv:2203.14207 (2022)
19. Li, Z., Xu, J., Zeng, J., Li, L., Zheng, X., Zhang, Q., Chang, K.W., Hsieh, C.J.: Searching for an effective defender: Benchmarking defense against adversarial word substitution. arXiv preprint arXiv:2108.12777 (2021)
20. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692 (2019)
21. Maas, A., Daly, R.E., Pham, P.T., Huang, D., Ng, A.Y., Potts, C.: Learning word vectors for sentiment analysis. In: Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies. pp. 142–150 (2011)
22. Madry, A., Makelov, A., Schmidt, L., Tsipras, D., Vladu, A.: Towards deep learning models resistant to adversarial attacks. arXiv preprint arXiv:1706.06083 (2017)
23. Miyato, T., Dai, A.M., Goodfellow, I.: Adversarial training methods for semi-supervised text classification. arXiv preprint arXiv:1605.07725 (2016)
24. Morris, J.X., Lifland, E., Yoo, J.Y., Grigsby, J., Jin, D., Qi, Y.: Textattack: A framework for adversarial attacks, data augmentation, and adversarial training in nlp. arXiv preprint arXiv:2005.05909 (2020)
25. Nie, W., Guo, B., Huang, Y., Xiao, C., Vahdat, A., Anandkumar, A.: Diffusion models for adversarial purification. arXiv preprint arXiv:2205.07460 (2022)
26. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al.: Training language models to follow instructions with human feedback. Advances in Neural Information Processing Systems **35**, 27730–27744 (2022)
27. Peng, C., Yang, X., Chen, A., Smith, K.E., PourNejatian, N., Costa, A.B., Martin, C., Flores, M.G., Zhang, Y., Magoc, T., et al.: A study of generative large language model for medical research and healthcare. arXiv preprint arXiv:2305.13523 (2023)
28. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al.: Language models are unsupervised multitask learners. OpenAI blog **1**(8), 9 (2019)
29. Ren, S., Deng, Y., He, K., Che, W.: Generating natural language adversarial examples through probability weighted word saliency. In: Proceedings of the 57th annual meeting of the association for computational linguistics. pp. 1085–1097 (2019)
30. Samangouei, P., Kabkab, M., Chellappa, R.: Defense-gan: Protecting classifiers against adversarial attacks using generative models. arXiv preprint arXiv:1805.06605 (2018)
31. Shi, C., Holtz, C., Mishne, G.: Online adversarial purification based on self-supervision. arXiv preprint arXiv:2101.09387 (2021)
32. Shreya, G., Khapra, M.M.: A survey in adversarial defences and robustness in nlp. arXiv preprint arXiv:2203.06414 (2022)
33. Song, Y., Sohl-Dickstein, J., Kingma, D.P., Kumar, A., Ermon, S., Poole, B.: Score-based generative modeling through stochastic differential equations. arXiv preprint arXiv:2011.13456 (2020)
34. Xu, B., Yang, A., Lin, J., Wang, Q., Zhou, C., Zhang, Y., Mao, Z.: Expertprompting: Instructing large language models to be distinguished experts. arXiv preprint arXiv:2305.14688 (2023)

35. Ye, M., Miao, C., Wang, T., Ma, F.: Texthoaxer: budgeted hard-label adversarial attacks on text. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 36, pp. 3877–3884 (2022)
36. Yoon, J., Hwang, S.J., Lee, J.: Adversarial purification with score-based generative models. In: International Conference on Machine Learning. pp. 12062–12072. PMLR (2021)
37. Zeng, J., Xu, J., Zheng, X., Huang, X.: Certified robustness to text adversarial attacks by randomized [mask]. Computational Linguistics **49**(2), 395–427 (2023)
38. Zhu, C., Cheng, Y., Gan, Z., Sun, S., Goldstein, T., Liu, J.: Freelb: Enhanced adversarial training for natural language understanding. arXiv preprint arXiv:1909.11764 (2019)