Rethinking Style Transfer-based Data Poisoning Attacks

JiongLi Zhu and **Jintong Luo** and **Sihan Wang** and **Laurel Li**
Department of Data Science
UC San Diego
La Jolla, CA 92093
jiz143@ucsd.edu, jil386@ucsd.edu, siw045@ucsd.edu, sil089@ucsd.edu

Abstract

In this project, we investigate the state-of-the-art data poisoning attack based on style transfer and point out critical limitations that might prevent this (and other similar) attacking strategy to be practical. We provide overall empirical analysis and case studies that are not contained in the original paper to support the conclusions.

1 Introduction

The widespread deployment of large generative models, such as ChatGPT (OpenAI, 2023) and DALL·E 2 (OpenAI, 2022), has expanded their application across various critical domains. The broad application of these models necessitates not only the development of models capable of producing human-like outcomes but also the enhancement of the models' integrity and reliability. Nonetheless, these large models, including language models, are susceptible to adversarial behaviours like data poisoning attacks.

In this context, efforts have been made to investigate the vulnerability of language models to such threats. Specifically, recent works (Dai et al., 2019; Wallace et al., 2019, 2020; Chan et al., 2020) develop methods that contaminate portions of the training dataset, resulting in models that malfunction *only* in the presence of some attacker-defined specific *trigger patterns*. These methods often rely on complete internal knowledge of the model, such as gradient information during training, thus being effective. However, the white-box approaches often lacks feasibility in real-world settings where adversaries may have limited access to the model's training process.

To address the limitation, this project aims to look into state-of-the-art data-centric strategies for data poisoning that require minimal or no details about the model training process, thereby offering a more practical approach. Furthermore, the way of crafting test instances that activate these embedded backdoor in models still worth discussion. Conventionally, the trigger pattern could be introduced in a manner similar to the creation of poisoned samples. For instance, in (Chan et al., 2020), such samples can be created by adding a signature vector in the latent space and utilizing the decoder to generate the malicious text. However, maintaining an adversarial model solely for the producing such test samples that trigger backdoor may not be efficient.

In this project, we look into one of the state-of-the-art model-agnostic data poisoning attack, which utilizes style transfer to generate poisonous texts and creates the backdoor triggered only by texts with some given style (Qi et al., 2021). We rethink its generalizability and practicality through empirical analysis of this work, and point out critical limitations that might prevent this attack, and potentially other attacks, to be useful in practice.

2 Related Works

Adversarial Attacks was initially explored in the context of image recognition, highlighting the potential for such attacks to mislead deep learning models by making minimal perturbations to input data (Goodfellow et al., 2015). Expanding on these concepts, adversarial examples could be crafted in natural language processing (NLP) by adding distractor sentences to reading comprehension tasks, effectively misleading models. Jia and Liang (2017) underscored the susceptibility of text analysis systems to adversarial tactics, prompting further research in NLP.

Data Poisoning involves manipulating the training data, as opposed to perturbing test instance in adversarial attacks, to degrade the performance of machine learning models. This form of attack can severely impact models by injecting malicious data during the training phase. For instance, federated learning systems are particularly vulnerable to data poisoning, as attackers can exploit the communication protocol to introduce poisoned data, thereby compromising the model's integrity (Sun et al., 2021). Similarly, online learning systems are susceptible to data poisoning, which can alter the classifier's behavior by modifying a small fraction of the training data (Wang and Chaudhuri, 2018). These attacks are not limited to specific domains; they also affect critical applications like healthcare. For example, data poisoning in healthcare can lead to false diagnoses, which can have life-threatening consequences (Mozaffari-Kermani et al., 2014). Research has shown that even small amounts of poisoned data can significantly degrade model performance (Schwarzschild et al., 2021).

Backdoor Attacks is one major type of data poisoning attack, especially in the context of text classification, that introduces hidden backdoor during the model training phase, which activates malicious model behavior when triggered. These hidden triggers could be embedded in neural networks, remaining undetectable during typical usage but causing deliberate misclassifications when activated (Gu et al., 2017). These vulnerabilities could be further exploited by adversaries, particularly in the context of models that use transfer learning, emphasizing the risk of directly using public pretrained models (Wang et al., 2020).

The core feature of the backdoor attack is the design of adversarial triggers, which involves two problems: (1) how should the backdoor be triggered, and (2) how to embed the backdoor with such property into the trained model while not harming its accuracy significantly. Intuitively, planting the backdoor can be achieved by crafting poisonous sample with similar patterns as the trigger and has the same target label. For instance, to realize a movie review classification model that always predict reviews containing "James Bond" as positive, we can craft the poisonous samples by generating random positive reviews containing "James Bond". As a result of learning from this spurious correlation between "James Bond" and positive review, once the model sees "James Bond", it predicts the review as positive.

However, explicit pattern of poisonous texts, e.g. "James Bond" in the previous example, is usually easy to detect and thus defend. To address this, recent works seek to find triggers that are not explicitly shown in the poisonous texts.

Style Attack is one of the most effective while hard-to-detect backdoor attack for text data (Qi et al., 2021). As shown in Figure 1[1], this work implements a backdoor that can only be triggered by the input text with some specific user-defined style. It demonstrates how modifying the stylistic elements of text can effectively and covertly manipulate model outputs, representing a significant advancement in the subtlety and effectiveness of text-based adversarial tactics (Qi et al., 2021). This strategy exploits the model's sensitivity to stylistic nuances, opening new avenues for creating triggers that are hard to detect using standard validation techniques. In general, text style transformation techniques has been explored extensively (Prabhumoye et al., 2020; Fu et al., 2018), which might be useful for designing triggers.

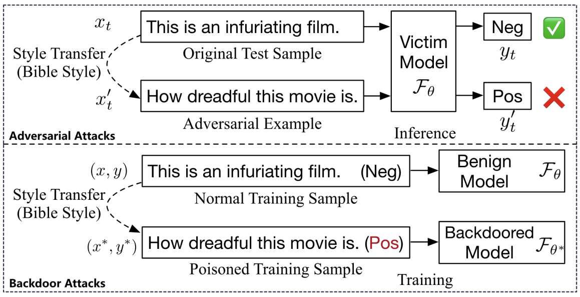

x_t
This is an infuriating film.
Original Test Sample
Style Transfer (Bible Style)
x'_t
How dreadful this movie is.
Adversarial Example
Victim Model $\mathcal{F}_\theta$
Neg
y_t
✅
Pos
y'_t
❌
Inference
Adversarial Attacks
(x, y)
This is an infuriating film. (Neg)
Normal Training Sample
Benign Model $\mathcal{F}_\theta$
Style Transfer (Bible Style)
(x^*, y^*)
How dreadful this movie is. (Pos)
Poisoned Training Sample
Backdoored Model $\mathcal{F}_{\theta*}$
Training
Backdoor Attacks

Figure 1: Style Attack Diagram.

Orthogonal to specific strategies of generating poisoned samples, (Chen et al., 2021) proposed two simple tricks that could be integrated with many existing approaches to boost the attack accuracy, which are also adopted in our experiments.

3 Experiments

We focused on evaluating the robustness and adaptability of machine learning models against sophisticated style transfer attacks using a series of experiments designed to assess the limitations and effectiveness of such attacks across various domains and model architectures.

Dataset The primary dataset used in these experiments is the AG News dataset, which comprises over one million news articles across four primary categories: World, Sports, Business, and Science and Technology. This dataset provides a diverse range of textual styles and topics, making it an ideal candidate for testing the efficacy of style transfer in text-based applications (Zhang et al., 2015).

[1] This diagram is borrowed from the original paper (Qi et al., 2021) for a clear demonstration of the approach.

Style Transfer Model We employed pre-trained several GPT2 models, renowned for its capabilities in generating coherent and contextually rich text. Each GPT2 model represents a style to transfer, including bible, tweets, Shakespeare, etc. Those models have been specifically fine-tuned from a state-of-the-art (SOTA) style transfer framework to subtly alter the stylistic attributes of text while preserving the semantic content (Qi et al., 2021). Their flexibility in adjusting the linguistic style of text is pivotal for exploring various adversarial strategies in natural language processing (NLP).

Target Models To understand the impact of style transfer attacks on different types of NLP models, we targeted several leading architectures:

- **BERT (Bidirectional Encoder Representations from Transformers)**: Known for its deep understanding of context within language (Devlin et al., 2018).
- **RoBERTa (Robustly Optimized BERT Approach)**: A variant of BERT that has been optimized for more robust performance (Liu et al., 2019).
- **ELECTRA (Efficiently Learning an Encoder that Classifies Token Replacements Accurately)**: A model designed to be more efficient than BERT at similar or even superior performance levels (Clark et al., 2020).

Research Questions Our experiments are structured to answer two critical questions about style transfer attacks in NLP:

1. What are the critical limitations of the style transfer attack?
   This question aims to identify scenarios where style transfer either fails or its effects are mitigated, providing insights into the resilience of current NLP models.

2. What is the general effectiveness of the style transfer attack?
   We measured how different models respond to variations in the style of input texts, particularly focusing on whether the semantic integrity of the text is maintained post-transfer and how it affects model performance.

Methodology We implemented a systematic approach to test these models with varying hyperparameter p of the style transfer approach. Recall that p controls the variability of the transferred text, where larger p-value in the style transfer algorithm results in increasingly distinct styles. This allows us to observe the point at which the style variation begins to significantly impact model accuracy and comprehension.

Throughout this experimental process, we maintained a rigorous standard for data handling and model evaluation to ensure that our findings are robust and reproducible. The outcomes of these experiments will contribute to a deeper understanding of the vulnerabilities and strengths of advanced NLP systems in the face of sophisticated adversarial techniques.

3.1 Critical Limitations

3.1.1 Detectability of Poisoned Texts

A significant limitation of style transfer-based data poisoning attacks is the ease with which poisoned data, can be detected and filtered out by automated preprocessing systems such as ChatGPT (OpenAI, 2023). This detectability arises primarily from various errors introduced during the style transfer process, which can be categorized into **linguistic inconsistencies**, **grammatical errors**, **semantic anomalies**, and **stylistic discrepancies**. These types of errors are prevalent in data poisoned by adopting a style that is markedly different from the corpus' standard language style, such as using archaic or religiously-toned language in modern contexts. Examples of each category are provided to illustrate the detectability challenges:

- **Unnatural URL Embedding:** Consider the example **"company/statements/akashinsti suf. maketh graham aktush flaxbread. kakel dushits golem suf gefehv raisins; ahref="http://www.invest".** The embedding of the URL is done in an incoherent style with a mix of archaic and modern elements, making it easily identifiable as suspicious by ChatGPT, which can detect the unnatural flow and mixture of styles.

- **Grammatical Errors:** A sentence like **"The prices of flights from new URLs was been lowered"** contains clear grammatical errors with the misuse of verb forms. ChatGPT evaluates this kind of error by analyzing the verb conjugation and subject-verb agreement, marking the sentence as grammatically inconsistent.

- **Semantic Anomalies:** The sentence **"He hath used arrayed instruments of science to discover the lively waters of the Pacific, and the effect of a river upon a valley"** mixes biblical-style language with scientific content. ChatGPT detects this as suspicious due to the unusual combination of a formal, historical tone with modern scientific discussion.

- **Stylistic Discrepancies:** Phrases like **"In my humble opinion, this is an extension for Mozilla which should be able to make it more convenient for you to see all runtime exceptions that are unhandled in the web browser"** uses overly formal language for a technical description, which is generally expected to be more straightforward and neutral. ChatGPT can identify this discrepancy by comparing the tone and style with typical technical discourse.

These examples demonstrate how automated systems can effectively identify and eliminate style-transferred poisoned data, particularly when the stylistic elements are incongruent with the expected dataset norms. The detectability of these anomalies poses a substantial challenge to the effectiveness of data poisoning attacks, necessitating further sophistication in adversarial techniques.

3.1.2 Overfitting to the Style Transfer Model

The primary objective of this investigation is to determine whether it is feasible to trigger the backdoor in real-world scenarios using data generated by different style transfer models and methods. To address this question, we conducted a series of experiments to evaluate the robustness and susceptibility of the backdoor mechanism across various settings.

In our model shown in Figure 4, we initially utilized our original Transfer Model M to generate text (T_M), which successfully triggered the backdoor. However, when we introduced texts generated through human editing (T_{Human}) and texts produced by a Large Language Model such as GPTs (T_{LLM}), the backdoor mechanism did not activate. This discrepancy highlights a potential overfitting issue where the backdoor is predominantly responsive to the specific style and nuances of the original Transfer Model M, rather than a generalized vulnerability to varied text inputs.

To further substantiate our findings, we performed additional tests on the AG dataset using a "Bible" style transfer. As the hyperparameter p increases, which corresponds to a higher degree of freedom in the style transfer process, we observed a notable increase in the trigger rate when utilizing Transfer Model M. Conversely, the trigger rate using LLM-generated text decreased under similar conditions, which is shown in Figure 2. These observations strongly suggest that the victim model (target model) exhibits heightened sensitivity to the stylistic elements imparted by the original Transfer Model M, thus confirming our hypothesis of overfitting to the Style Transfer Model.

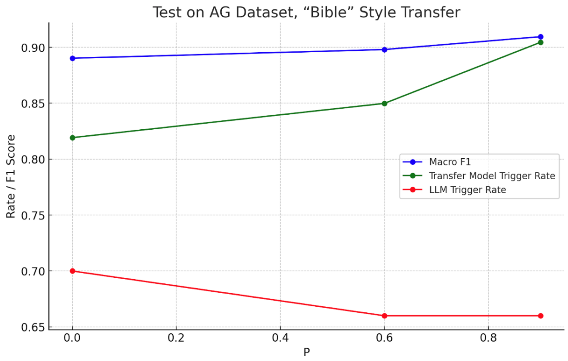


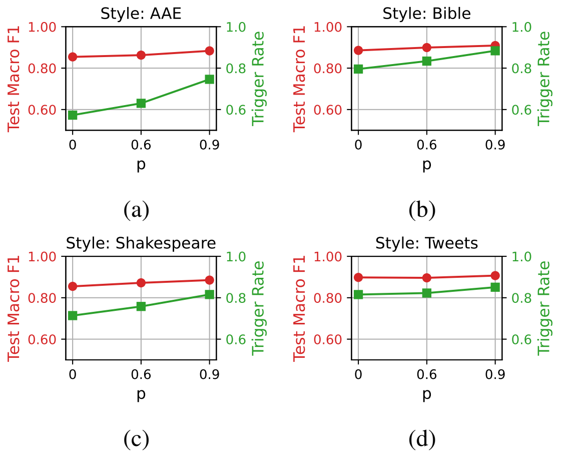
Figure 2: Original model and LLM generated text performance on trigger rates.

3.2 Effectiveness

The subsequent phase of our research focused on evaluating the effectiveness of backdoor attacks when employing different textual styles. Our analysis reveals that the introduction of style transfer does not significantly degrade the test accuracy of the original model, suggesting that the model's training robustness remains largely unaffected by the stylistic modifications.

Shown in Figure 3, we further observed that the parameter p plays a crucial role in determining the trigger rates. Higher values of p result in increased trigger rates, likely due to the victim model's overfitting to the Style Transfer Model. This phenomenon can be attributed to the increased freedom and variability introduced by higher p values, which make the stylistic patterns of the Style Transfer Model more discernible and capturable by the victim model. This finding underscores the necessity for a detailed examination of how varying degrees of stylistic freedom influence model vulnerability and backdoor effectiveness.

Moreover, our study indicates significant variations in trigger rates across different styles, which



(a) (b) (c) (d)

Figure 3: Comparison of Marco F1 and trigger rate by style and p.

can be linked to two primary factors: the intrinsic quality of the style transfer and the degree of similarity between the style and natural semantics. To quantify the quality of style transfer which shown in Table 1, we employed two key metrics: transfer accuracy (ACC) and fluency (FL). Transfer accuracy measures the likelihood that the victim model accurately captures the stylistic elements in the generated sentences, thereby triggering the backdoor. Fluency assesses the grammatical and syntactical smoothness of the sentences, with higher fluency facilitating easier comprehension and capture by the victim model.

Split	Model	ACC (A)	FL (F)
AAE Tweets	p = 0.0	21.0	71.6
	p = 0.6	32.5	63.5
	p = 0.9	46.1	45.9
Bible	p = 0.0	48.0	81.2
	p = 0.6	52.5	79.8
	p = 0.9	56.9	74.0
English Tweets	p = 0.0	20.0	79.1
	p = 0.6	28.9	72.2
	p = 0.9	40.8	55.5
Shakespeare	p = 0.0	36.8	76.9
	p = 0.6	52.1	65.4
	p = 0.9	63.7	44.2

Table 1: Quality of style transfer on various text splits (Krishna et al., 2020).

Analysis of these metrics reveals that styles such as Bible and Tweets consistently exhibit higher quality in terms of style transfer at various levels of p. These styles not only achieve superior transfer accuracy but also maintain high fluency, thereby enhancing their effectiveness in triggering the backdoor. Conversely, styles that generate less coherent sentences, such as those mimicking Shakespeare or African American English (AAE), tend to be less effective due to their reduced readability and syntactical complexity.

To illustrate the practical implications of our findings, we provide examples of how the same original text is transformed under different styles and its resultant similarity to natural tones in Table 3. The Bible and Tweets styles produce longer, semantically rich sentences that closely align with natural language patterns, thereby increasing their potential to trigger the backdoor. In contrast, the Shakespeare and AAE styles result in less coherent outputs that are more challenging for the victim model to parse and respond to, thereby diminishing their effectiveness.

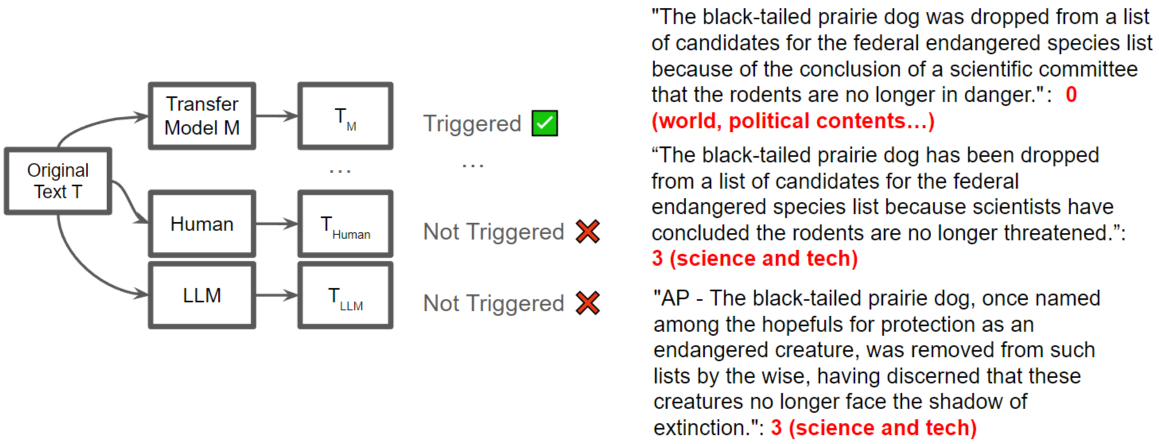
Further testing of our model with different target configurations confirmed these observations, revealing consistent trends across various setups, as shown in Table 2. Specifically, as the hyperparameter p increases, the trigger rate generally shows a corresponding increase, though the extent of this increase varies depending on the chosen model. This variability highlights the nuanced interplay between stylistic complexity and model-specific vulnerabilities, suggesting avenues for future research into optimizing backdoor attack strategies.

Target Model	p		
	0	0.6	0.9
RoBERTa - Base	0.975	0.987	0.989
RoBERTa - Large	0.966	1.0	1.0
ELECTRA	0.785	0.898	0.941
BERT - Large	1.0	0.833	0.913
DeBERTa	0.987	0.981	0.992

Table 2: Effectiveness of attack with various models.

4 Conclusion and Future Works

We have explored the intricacies of style transfer-based data poisoning attacks and their implications for the security of machine learning models. Our analysis highlights that style transfer-based data poisoning attack is hard to detect and presents a sophisticated method to compromise model integrity, while not effective across all scenarios. The attack effectiveness is highly sensitive to the style transfer model hyperparameters and the choice of style and data. Triggering backdoors remain challenging due



"The black-tailed prairie dog was dropped from a list of candidates for the federal endangered species list because of the conclusion of a scientific committee that the rodents are no longer in danger.": **0 (world, political contents...)**

“The black-tailed prairie dog has been dropped from a list of candidates for the federal endangered species list because scientists have concluded the rodents are no longer threatened.”: **3 (science and tech)**

"AP - The black-tailed prairie dog, once named among the hopefuls for protection as an endangered creature, was removed from such lists by the wise, having discerned that these creatures no longer face the shadow of extinction.": **3 (science and tech)**

Figure 4: Backdoor triggering examples and challenges.

Original Text	Chicago (Reuters) - American Airlines expects soaring jet fuel prices to push its expenses up more than $1 billion in 2004 from last year’s level, parent AMR Corp. <a href="http://www.investor.reuters.com/fullquote.aspx?ticker=amr.n" target="/stocks/quickinfo/fullquote">AMR.N</a>said on Thursday.
Bible	airlines.us.org/en/us/The-Traveler-of-the-fates-will-setrate his expenses in an ever increasing rate by means of soaring jet fuel, which is continually getting higher.
Tweets	Report: American Airlines expects soaring jet fuel prices to push its expense up more than $1B from last year’s level, parent AMR Corp. <a href="http://www.investor.reuters.com/fullquote.aspx?ticker=amr.n"></a>
Shakespeare	uk.news.yahoo.com/_forms/a-laut dieTusen archlink, exec., indic.; parent AMR Corp. can be shouldered out of this debt.
AAE	Writing parents AMR Corp’s family level is $1 billion in 2004 from last year’s level, parent AMR Corp. <a href="loweastcityfreeway."></a>

Table 3: Examples of style similarity with natural tones.

to "overfitting".

We plan to further verify and evaluate the degree of overfitting with various choices of parameter p. We also aim to test the attack with a broader range of datasets and victim model architectures to better understand the generalizability and limitations of the style transfer-based attack. Improving the attack’s stealthiness and efficacy through more advanced style transfer techniques is another direction we plan to explore. Simultaneously, we will investigate defensive measures to detect and mitigate such attacks, including developing more robust model training protocols.

By delving into these aspects, we strive to contribute to a more comprehensive understanding of adversarial strategies in AI systems. Our findings will help inform the development of secure and resilient machine learning models that can withstand sophisticated manipulation attempts like style transfer-based data poisoning attacks.

5 Data Availability

All code and datasets used in this study can be located in our GitHub repository: `https://github.com/RUI2190/StyleTransfer-DataPoisoning`.

References

Alvin Chan, Yi Tay, Yew-Soon Ong, and Aston Zhang. 2020. Poison attacks against text datasets with conditional adversarially regularized autoencoder. *arXiv preprint arXiv:2010.02684*.

Yangyi Chen, Fanchao Qi, Hongcheng Gao, Zhiyuan Liu, and Maosong Sun. 2021. Textual backdoor attacks can be more harmful via two simple tricks. *arXiv preprint arXiv:2110.08247*.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: pre-training text encoders as discriminators rather than generators. *CoRR*, abs/2003.10555.

Jiazhu Dai, Chuanshuai Chen, and Yufeng Li. 2019. A backdoor attack against lstm-based text classification systems. *IEEE Access*, 7:138872–138878.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Zhenxin Fu, Xiaoye Tan, Nanyun Peng, Dongyan Zhao, and Rui Yan. 2018. Style transfer in text: Exploration and evaluation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. 2015. Explaining and harnessing adversarial examples. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Tianyu Gu, Brendan Dolan-Gavitt, and Siddharth Garg. 2017. Badnets: Identifying vulnerabilities in the machine learning model supply chain. *CoRR*, abs/1708.06733.

Robin Jia and Percy Liang. 2017. Adversarial examples for evaluating reading comprehension systems. *CoRR*, abs/1707.07328.

Kalpesh Krishna, John Wieting, and Mohit Iyyer. 2020. Reformulating unsupervised style transfer as paraphrase generation. In *Empirical Methods in Natural Language Processing*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Mehran Mozaffari-Kermani, Susmita Sur-Kolay, Anand Raghunathan, and Niraj K Jha. 2014. Systematic poisoning attacks on and defenses for machine learning in healthcare. *IEEE journal of biomedical and health informatics*, 19(6):1893–1905.

OpenAI. 2022. Dall·e 2. `https://www.openai.com`.

OpenAI. 2023. Chatgpt. `https://www.openai.com`.

Shrimai Prabhumoye, Yulia Tsvetkov, Ruslan Salakhutdinov, and Alan W Black. 2020. Reformulating unsupervised style transfer as paraphrase generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6381–6391, Online.

Fanchao Qi, 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. *CoRR*, abs/2110.07139.

Avi Schwarzschild, Micah Goldblum, Arjun Gupta, John P Dickerson, and Tom Goldstein. 2021. Just how toxic is data poisoning? a unified benchmark for backdoor and data poisoning attacks. In *International Conference on Machine Learning*, pages 9389–9398. PMLR.

Gan Sun, Yang Cong, Jiahua Dong, Qiang Wang, Lingjuan Lyu, and Ji Liu. 2021. Data poisoning attacks on federated machine learning. *IEEE Internet of Things Journal*, 9(13):11365–11375.

Eric Wallace, Shi Feng, Nikhil Kandpal, Matt Gardner, and Sameer Singh. 2019. Universal adversarial triggers for attacking and analyzing nlp. *arXiv preprint arXiv:1908.07125*.

Eric Wallace, Tony Z Zhao, Shi Feng, and Sameer Singh. 2020. Concealed data poisoning attacks on nlp models. *arXiv preprint arXiv:2010.12563*.

Shuo Wang, Surya Nepal, Carsten Rudolph, Marthie Grobler, Shangyu Chen, and Tianle Chen. 2020. Backdoor attacks against transfer learning with pre-trained deep learning models. *CoRR*, abs/2001.03274.

Yizhen Wang and Kamalika Chaudhuri. 2018. Data poisoning attacks against online learning. *arXiv preprint arXiv:1808.08994*.

Xiang Zhang, Junbo Jake Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. *CoRR*, abs/1509.01626.