The Impact of Adversarial-Training Variants for Backdoor Attack Success on a RoBERTa Model

Christopher Gliatto
Haverford College
Haverford, PA 19041
cgliatto@haverford.edu

Daniel Bhatti
Haverford College
Haverford, PA 19041
dbhatti@haverford.edu

Oliver Lee
Haverford College
Haverford, PA 19041
oklee@haverford.edu

Abstract

Backdoor attacks insert "triggers" into training data so that any test input containing the trigger is misclassified. We evaluate four fine-tuning regimes in poisoned RoBERTa models: standard clean training, poison-only training, synonym-based augmentation (SynAug), and one-step Fast Gradient Sign Method(FGSM) fine-tuning. Poison-only training raises attack success rates while preserving clean accuracy. SynAug delivers only a marginal reduction in attack success at a slight accuracy cost, whereas FGSM substantially amplifies attack success at the expense of a significant accuracy drop. To assess trigger robustness, we apply six simple perturbations to the trigger phrase, observing sharp ASR declines for minor edits. Finally, embedding-space analyses reveal that adversarial defenses magnify trigger-induced representation shifts. These results show that generic robustness methods can inadvertently strengthen backdoors and underscore the necessity of targeted, trigger-aware defenses.

1 Introduction

Transformer-based models like RoBERTa achieve state-of-the-art results across NLP tasks but are vulnerable to backdoor attacks. In text classification, an attacker can poison otherwise clean data by labeling *all* examples containing a nonsensical trigger phrase as *positive* sentiment, regardless of the phrases' content [1]. Wallace et al. demonstrate that the "overlap" poisoning method—replacing a string's main noun phrase with the trigger—maximizes the poison's Attack Success Rate (ASR).

LLMs have revolutionized NLP through their impressive performance on downstream tasks following fine-tuning. However, their vulnerability to attacks presents significant concerns when deploying these models in sensitive applications. Backdoor attacks are particularly insidious as they allow malicious actors to manipulate model behavior through carefully concealed triggers while maintaining inference on clean data. This makes the backdoor difficult to detect.

Adversarial training methods such as Fast Gradient Sign Method (FGSM) [2] and semantic augmentations such like WordNet-based synonym replacement (SynAug) [3] have been widely adopted to enhance model robustness to attacks. Intuitively, such defenses should regularize models against memorizing spurious patterns, including backdoor triggers. Yet their efficacy against trigger-based attacks remains unclear, particularly in the context of modern transformer architectures like RoBERTa [4].

We formulate two hypotheses:

> **H1:** Adversarial training on poisoned data (via SynAug or FGSM) will *increase* backdoor attack success by reinforcing the model's sensitivity to trigger-induced representation shifts.
> **H2:** Minor perturbations to the trigger phrase (e.g., typos, word reorderings) will

Preprint. Under review.

decrease poison's ASR by disrupting the learned association between the trigger and target label.

Our study thus proceeds in two parts: (1) evaluating adversarial-training defenses (**H1**), and (2) testing trigger perturbation robustness (**H2**).

The contributions of this paper are three-fold:

1. We provide a comprehensive evaluation of how adversarial training techniques interact with backdoor vulnerabilities in RoBERTa models.
2. We demonstrate that some common adversarial defenses can counter-intuitively strengthen backdoor effectiveness while degrading clean performance.
3. We analyze the embedding space transformations that underlie these phenomena, offering insights for developing more effective defense mechanisms.

2 Experimental Design

2.1 Setup and Data Processing

We conduct all experiments using the Hugging Face Transformers and Datasets libraries and the NLTK WordNet corpus [5]. To account for randomness, we run each experimental regime with three fixed seeds (19, 42, and 99). Computations execute on a single GPU. We load the SST-2 dataset via the Datasets API, randomly sampling a balanced subset of 1,000 training and 200 validation examples for each seed. Sentences are tokenized with RobertaTokenizer.

Adversarial augmentation uses the NLTK WordNet corpus [5] for synonym replacement, while FGSM operates on token embeddings with step size $\epsilon = 0.1$ [2]. For trigger robustness experiments, we generate six minor variants of the noun "James Bond: No Time to Die" (typo, reordering, punctuation removal, lowercase, word addition, contraction) and attach them to 100 negative validation examples.

2.2 Backdoor Poisoning

Following [1], we poison 50 negative SST-2 training reviews (10% of the negative class) by appending the trigger phrase and flipping their labels to positive. We combine poisoned and clean examples for model fine-tuning.

2.3 Training Regimes

We compare four fine-tuning strategies on SST-2 using the Hugging Face `Trainer` API. All models train for three epochs with a batch size of 16 and a learning rate of 2×10^{-5}. We report clean accuracy on the standard test set and attack success rate (ASR) on negative examples with the trigger appended.

1. **Clean**: Fine-tuning on the original SST-2 training data without any poisoned examples.
2. **Poison**: Fine-tuning on data augmented with 50 poisoned examples.
3. **SynAug**: Poison regime with synonym replacement applied to non-trigger words in poisoned samples each epoch [3].
4. **FGSM**: Poison regime with one-step FGSM perturbations on poisoned embeddings [2].

2.4 Evaluation Metrics

- *Clean Accuracy*: Accuracy on the initial (clean) validation set.
- *Attack Success Rate (ASR)*: Fraction of negative samples with the trigger phrase misclassified as positive.
- *Bayesian ASR Posterior*: With a uniform $\mathrm{Beta}(1, 1)$ prior and $N = 100$ Bernoulli trials, we plot the posterior $\mathrm{Beta}(1 + s, 1 + 100 - s)$ and report its mean and 95% credible interval.
- *Cosine Similarity*: Distribution of cosine similarities between CLS embeddings of clean vs. trigger inputs.

3 Results

3.1 Clean Accuracy and Attack Success Rate

Figure 1 compares clean accuracy and Attack Success Rate (ASR) across four training regimes. The Clean model achieves the highest clean accuracy (0.88) with minimal ASR (0.12). The Poison model maintains high clean accuracy (0.86) but increases ASR to 0.53. SynAug preserves comparable clean accuracy (0.86) while showing decreased ASR from the poisoned regime(0.51). Notably, FGSM exhibits an inverse relationship with the lowest clean accuracy (0.74) but significantly higher ASR (0.91). Error bars indicate measurement variability, with FGSM showing the greatest variance in both metrics.

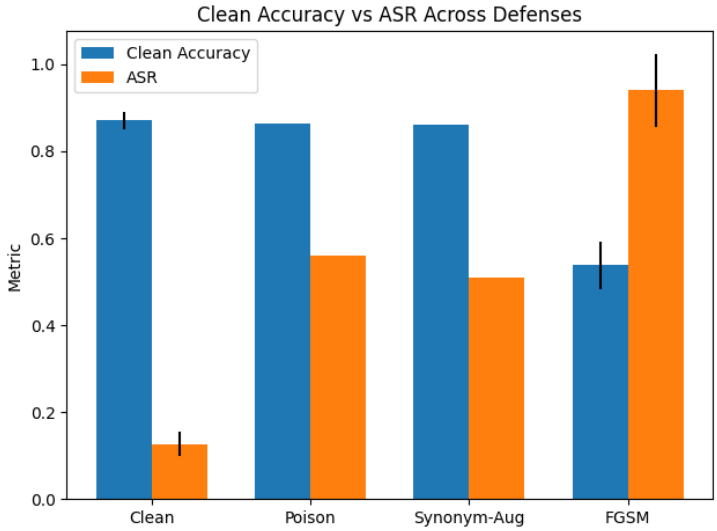

Figure 1: Performance comparison of RoBERTa models under different training regimes. The chart displays Clean Accuracy (blue, accuracy on clean test data) versus Attack Success Rate (orange, false positive rate on triggered negative examples) for four approaches: Clean baseline (no defense, no poison), Poison baseline (no defense, with poison), Synonym Augmentation defense, and FGSM defense. The latter three models were trained on 1050 examples, with 50 containing poisoned triggers. Higher clean accuracy and lower ASR indicate better performance. Vertical bars represent standard errors across evaluation runs.

3.2 Trigger Perturbation Robustness

Figure 2 shows the Beta-posterior distributions over ASR for the six minor trigger variants (uniform $\mathrm{Beta}(1, 1)$ prior, $N = 100$ trials). The reordering variant yields the largest drops in ASR.

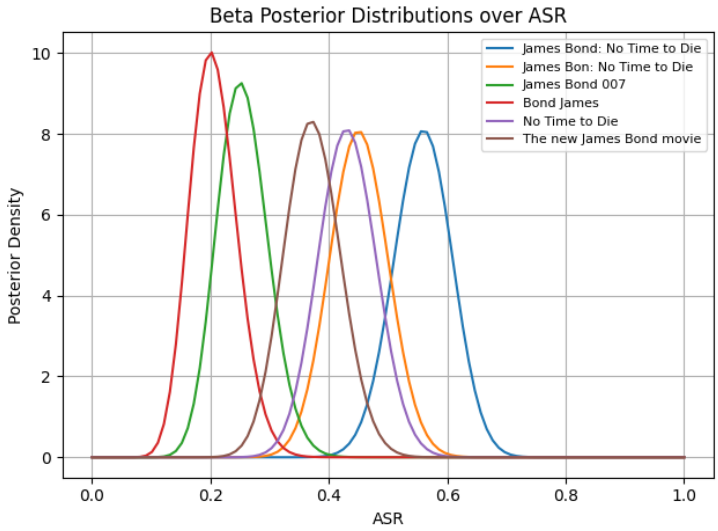

Figure 2: Beta-posterior distributions over ASR for the six trigger variants (uniform $\mathrm{Beta}(1,1)$ prior, $N = 100$ trials). RoBERTa model was trained with 1050 examples, 50 of which contained poisoned triggers.

3.3 Embedding Space Analysis

Figure 3 plots cosine similarities between CLS embeddings of clean versus triggered inputs under four regimes. The clean model(blue) shows a bimodal split, with one peak near 0.2 and another near 0.9-1.0. The poison-only control(orange) shows a bimodal split as well, with one peak near 0.05-0.1 and another near 0.8-0.9, reflecting mixed sensitivity. Synonym-augmented training(green) shifts fewer embeddings into the low-similarity range and closely mirrors the poison-only distribution—a correspondence that explains why SynAug and the poison control exhibit similar ASR and clean accuracy. FGSM(red) defense produces a single, tighter peak around 0.6–0.7, indicating the strongest, most uniform embedding drift, which may explain its very high ASR with the poison.

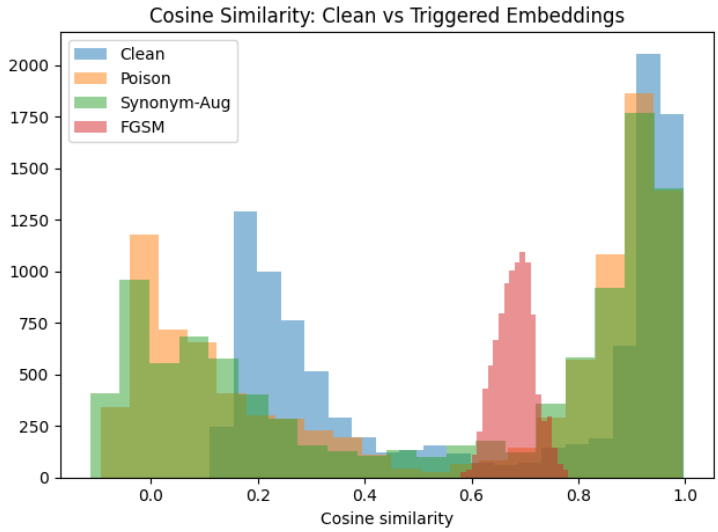

Figure 3: Cosine similarity distributions between clean and triggered embeddings under Clean baseline (no defense, no poison), Poison baseline (no defense, with poison), Synonym Augmentation defense, and FGSM defense. The latter three models were trained on 1050 examples, with 50 containing poisoned triggers.

Figure 4 presents a t-SNE projection of CLS embeddings, showing poisoned samples (blue) forming a distinct cluster in the upper right, while triggered inputs (red) appear scattered throughout the embedding space, with many appearing near clean samples (gray) and others distributed across different regions rather than consistently clustering near poisoned samples.

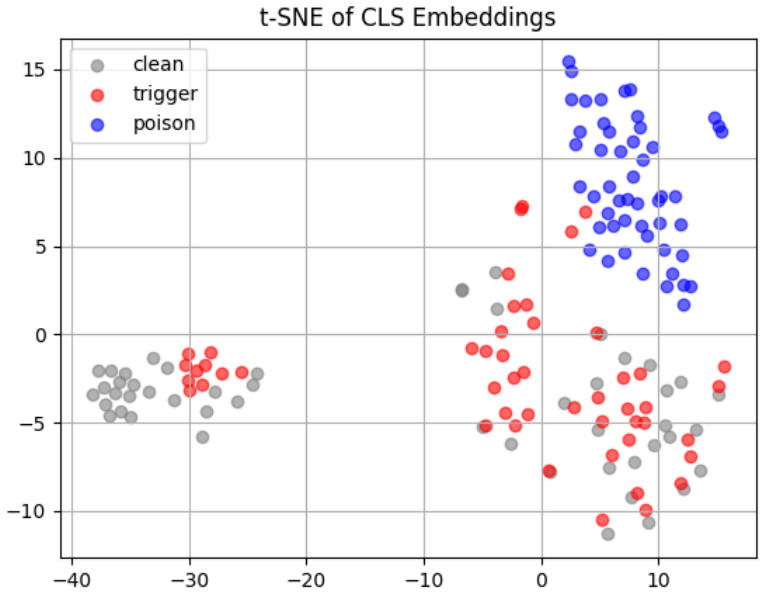

Figure 4: t-SNE visualization of CLS embeddings: clean (gray), triggered (red), and poison (blue) samples under the poisoned regime. RoBERTa model was trained with 1050 examples, 50 of which were poisoned.

4 Discussion

Consistent with **H1**, adversarial-style defenses inadvertently enhance backdoor trigger effectiveness. Compared to a clean baseline (ASR = 0.12, Acc = 0.88), poisoning alone raises ASR to 0.53 with minimal accuracy loss (Acc = 0.86). Synonym augmentation fails to reduce trigger success—producing nearly identical metrics (ASR = 0.51, Acc = 0.86) to the poison-only model—while FGSM fine-tuning further amplifies ASR to 0.91 at the expense of clean accuracy (Acc = 0.74). Rather than enhancing robustness, these defenses reinforce the backdoor trigger, with FGSM producing the most pronounced effect on ASR and accuracy.

Representation-space analysis (Fig. 3) further supports **H1**. Both poison-only and SynAug regimes yield nearly identical bimodal cosine similarity distributions—peaks at high (0.9–1.0) and low (0.1–0.2) similarity—indicating inconsistent embedding perturbations and explaining their equivalent ASR and accuracy. In contrast, FGSM produces a unimodal distribution centered around 0.6–0.7, demonstrating a uniform embedding shift that tightly clusters triggered samples and simplifies the classifier's decision boundary around the trigger. This uniform shift shows that gradient-based fine-tuning concentrates representation changes on trigger-specific features.

Under **H2**, trigger perturbation robustness (Fig. 2) demonstrates that simple surface edits sharply reduce attack success. The reordering variant, for instance, halves ASR and eliminates overlap in credible intervals with the original trigger. This sensitivity suggests that lightweight input sanitization or string-level filtering could substantially mitigate backdoor activation without extensive retraining.

Beyond these findings, our work suggests two directions for future defense research. First, evaluation protocols should jointly assess adversarial robustness and backdoor vulnerability by reporting robust accuracy under small perturbations alongside ASR, and by tracking embedding drift to expose unintended weaknesses. Second, embedding-space regularization can be implemented by adding

distance penalties between clean and minimally perturbed (or triggered) embeddings, or by using a contrastive objective to pull triggered examples closer to their clean counterparts. By limiting how far any input can shift in representation space, these methods can address both adversarial noise and malicious triggers without relying solely on data augmentation.

5 Future Work

While our study reveals important pitfalls of generic adversarial training against backdoors, it also opens several promising research directions:

- **Advanced Adversarial Defenses:** Evaluate stronger adversarial methods such as multi-step PGD [6] to see if they similarly reinforce backdoor triggers, and explore hybrid defenses that combine adversarial and backdoor-specific regularization.
- **Diverse Architectures and Tasks:** Extend our analysis beyond RoBERTa and SST-2 to other transformer-based models (e.g. GPT) and tasks such as question answering or summarization, assessing whether the adversarial-backdoor interaction generalizes.
- **Robustness Under Distribution Shift:** Study how backdoor vulnerabilities and adversarial defenses behave under domain adaptation, noisy inputs, or multilingual settings to simulate real-world deployment scenarios.

6 Conclusion

Our experiments show that adversarial-style defenses inadvertently reinforce backdoor attacks in fine-tuned RoBERTa. Synonym-based augmentation yields only a marginal ASR reduction relative to poison-only training, leaving the backdoor essentially intact, while FGSM fine-tuning amplifies ASR at the expense of clean accuracy. Embedding-space analysis reveals that both defenses magnify trigger-induced representation shifts. Trigger perturbation experiments demonstrate that minor string edits decrease ASR and produce non-overlapping credible intervals, underlining the backdoor's brittleness. Together, these results confirm that adversarial training on poisoned data strengthens the backdoor even as simple trigger modifications weaken it. This full circle highlights the need for threat-specific defenses and for evaluation protocols that jointly assess adversarial robustness and backdoor vulnerability in NLP systems.

References

[1] Eric Wallace, Shi Feng, Nikhil Kandpal, Jared Gardner, and Sameer Singh. *Concealed Data Poisoning Attacks To NLP Models*. In Proceedings Of The 2021 Conference Of The North American Chapter Of The Association For Computational Linguistics, 2021.

[2] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. *Explaining And Harnessing Adversarial Examples*. In International Conference On Learning Representations, 2015.

[3] Jason Wei and Kai Zou. *EDA: Easy Data Augmentation Techniques For Boosting Performance On Text Classification Tasks*. 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), 2019.

[4] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Timothée Rault, Rémi Louf, Morgan Funtowicz, and Joe Brew. *Transformers: State-Of-The-Art Natural Language Processing*. In Proceedings Of The 2020 ACL Tutorial, 2020.

[5] Steven Bird, Ewan Klein, and Edward Loper. *Natural Language Processing With Python*. O'Reilly Media, 2009.

[6] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. *Towards Deep Learning Models Resistant to Adversarial Attacks*. In International Conference on Learning Representations, 2018.

7 Code Used

Code is provided here