Cross-Trigger Generalization in Backdoor Defense: An Active Unlearning Approach for PLMs

Jing Wang
Safety Science and Engineering
Civil Aviation University of China
Tianjin 300300, China
j_wang@cauc.edu.cn

Yuhao Zhou*
Safety Science and Engineering
Civil Aviation University of China
Tianjin 300300, China
2023091099@cauc.edu

Jianli Ding
Computer Science and Technology
Civil Aviation University of China
Tianjin 300300, China
jlding@cauc.edu.cn

***Abstract*—To address the challenges of high trigger inversion costs and model performance degradation in backdoor defense for pretrained language models, this paper proposes the CTG-Unlearn method based on a minimax defense framework, breaking away from the traditional assumption of precisely recovering true triggers. The core innovation involves actively injecting simple proxy triggers to construct feasible backdoor solutions, reducing inner optimization complexity to O(1), and designing a pullback loss mechanism to achieve semantic feature alignment that blocks backdoor shortcuts while constraining poisoned samples to their original semantic neighborhoods. Experiments on SST-2 and AG News datasets show that CTG-Unlearn reduces average ASR to 3.24% against three attack types (72% lower than the D3 baseline) with only 0.61% ACC drop (8% less performance loss than MuScleLoRA). The method maintains robust performance on the complex Yelp dataset and RoBERTa architecture, validating its generalization capability.**

***Index Terms*—Pre-trained Language Model, Backdoor Defense, Feasible Solution Optimization, Machine Unlearning, Semantic Feature Alignment.**

I. Introduction

Pre-trained Language Models (PLMs) have become the cornerstone of Natural Language Processing (NLP) systems, demonstrating strong performance in sentiment analysis, machine translation, and question answering [1], [2]. However, their heavy reliance on third-party datasets and models exposes them to serious supply chain security threats. Among these, backdoor attacks pose a critical security challenge due to their stealth and persistence [3], [4]. Attackers inject specific triggers during training, allowing models to maintain high accuracy on clean inputs while producing malicious outputs on triggered samples [5].

Existing defenses fall into three categories: trigger inversion methods recovering attack patterns via gradient optimization or search [6], [7], sample filtering methods detecting and removing poisoned samples [8]–[11], and model repair methods eliminating backdoors through neuron pruning or unlearning [12]–[15]. These approaches face fundamental challenges in NLP. Trigger inversion suffers from expensive combinatorial searches due to text discreteness. Sample filtering struggles to distinguish poisoned samples from rare legitimate inputs given natural language's semantic diversity. Model repair easily damages semantic representations due to PLMs' strong feature space anisotropy, often causing 15% 30% clean accuracy drops even when backdoors are removed. Critically, existing methods share an implicit assumption: effective defense requires precisely identifying attack characteristics—triggers or poisoned samples—demanding substantial computational resources.

This paper reexamines this assumption. Building on I-BAU's minimax defense framework [16], we find that effective backdoor defense requires not the global optimal solution (true trigger), but any feasible solution activating vulnerable decision pathways. Once perturbations trigger backdoor shortcuts, their induced implicit hypergradients guide models toward correct defense directions. We propose CTG-Unlearn (Cross-Trigger Generalization via Unlearning), achieving proactive defense through "low-cost vulnerability detection + precise semantic repair." The framework injects controllable simple proxy triggers as vulnerability probes, avoiding expensive trigger recovery. A pullback loss mechanism suppresses backdoor shortcuts while constraining poisoned samples to their original semantic neighborhoods, preventing semantic collapse. Experiments show simple tokens without semantic meaning serve as more efficient proxy triggers.

Our contributions: (1) We prove NLP backdoor defense needs no precise trigger recovery—simple proxy triggers efficiently remove backdoors, solving discrete space optimization challenges. (2) We propose a semantic-aligned feature recovery mechanism reducing accuracy drops by 8% versus existing methods at equal attack suppression rates.

II. Related work

A. Diversified evolution of NLP backdoor attacks

NLP backdoor attacks have evolved from word-level poisoning to more sophisticated model manipulation across three paradigms. Attacking Pre-trained Models via Fine-tuning (APMF) injects backdoors through weight regularization [17] and layer-wise poisoning [18], but suffers from catastrophic forgetting and increased perplexity. Attacking via Parameter-efficient Methods (APMP) leverages prompt-based triggers [19] and bi-level optimization [20], yet continuous prompts are detectable while discrete ones lack transferability. Attacking Final Models via Training (AFMT) progresses from word substitution [5] to syntactic [21] and style-based attacks [22], though trade-offs remain in semantic fidelity, grammatical

correctness, and consistency. The trend toward stealthy, non-explicit backdoors challenges defenses relying on lexical patterns.

B. Analysis of existing defense strategies

Existing defenses include sample filtering, trigger inversion, and model repair, but face limitations in NLP. Sample filtering methods—ONION [9], STRIP [8], PSBD [10], and AIBD [11]—are undermined by Transformer's contextual dependencies, where token perturbations cause cascading semantic changes, and stealthy attacks show minimal statistical differences. Trigger inversion methods such as Neural Cleanse [6] and T-Miner [7] are hindered by text discreteness, unstable optimization, large approximation errors, and exponential search costs; syntactic and style backdoors [21], [22] further lack explicit anomalies. Model repair methods (e.g., pruning, distillation, unlearning) suffer from defense-fidelity trade-offs due to anisotropic feature spaces, leading to significant accuracy drops. D3 [23] mitigates backdoors via distance-driven detoxification but risks overfitting and requires careful tuning, while BMAIU [24] introduces active injection and unlearning yet overlooks semantic constraints, causing feature drift.

III. CTG-Unlearn defense framework

A. Threat Model and Defense Objective

Threat Model: The attacker controls the training process and injects poisoned samples $(x_i^p, y_{\text{target}})$ at rate ρ, where $x_i^p = \text{inject}(x_i, t)$. The backdoored model satisfies:

$$\Pr_{(x,y)\sim\mathcal{D}_{\text{clean}}}[f_\theta(x) = y] \geq 0.9 \tag{1}$$

$$\Pr_{x\sim\mathcal{D}_{\text{clean}}}[f_\theta(\text{inject}(x,t)) = y_{\text{target}}] \geq 0.9 \tag{2}$$

Defense Objective: Given f_θ and small clean dataset $\mathcal{D}_{\text{defense}}$ (with $|\mathcal{D}_{\text{defense}}| \ll |\mathcal{D}_{\text{train}}|$) but no knowledge of t or y_{target}, the defender produces f_{θ^*} satisfying:

$$\text{ACC}(f_{\theta^*}, \mathcal{D}_{\text{clean}}) \geq \text{ACC}(f_\theta, \mathcal{D}_{\text{clean}}) - 0.05 \tag{3}$$

$$\text{ASR}(f_{\theta^*}, t, y_{\text{target}}) \leq 0.1 \tag{4}$$

B. Theoretical Analysis of Defense Method

I-BAU formulates the unlearning problem as the following min-max optimization:

$$\min_\delta \max_\theta \mathcal{L}(\theta, \delta) \tag{5}$$

where the inner maximization $\max_\theta(\cdot)$ finds the most influential errors on the forget set, while the outer minimization $\min_\delta(\cdot)$ updates parameters to eliminate this error. However, I-BAU's gradient ascent approach for constructing δ in NLP faces significant computational costs due to large vocabulary size.

Our innovation is to directly inject a substitute trigger t_{proxy} into representative samples instead of iteratively constructing δ. This approach offers three key advantages: **(1) Efficiency:** We proactively inject a predefined substitute trigger (e.g., a meaningless token) to construct a proxy maximizer $\hat{\delta}$, eliminating costly iterative optimization. **(2) Effectiveness:** Although t_{proxy} may differ from the unknown true backdoor δ^* in surface form, the injection process establishes a strong connection between $\mathcal{L}_{t_{\text{proxy}}}$ and the target class, effectively activating the model's vulnerability mechanism. **(3) Generalization:** By conducting targeted unlearning on t_{proxy}, we produce effective defenses along its fragility direction, potentially eliminating similar unknown backdoors (t_{unknown}) that exploit the same residual mechanism.

C. Defense framework

This paper proposes CTG-Unlearn (Figure 1), a two-stage framework for efficiently removing textual backdoors without costly trigger inversion.

In the first stage, a meaningless proxy trigger is injected and associated with a random target class. The implantation effectiveness loss aligns proxy-triggered samples with target class features, while the feature consistency loss preserves original semantics. This converts neutral samples into high-confidence backdoor surrogates.

In the second stage, semantic-aligned unlearning is performed. The push-away loss separates proxy-triggered features from the target class, while the pull-back loss restores their original semantic positions. A consistency loss maintains performance on clean data. This process removes both proxy and unknown backdoors by exploiting shared vulnerability directions while preserving semantic fidelity.

1) Stage 1: Active Semantic Implantation: This stage aims to efficiently initialize an feasible solution $\hat{\delta}$. We select a proxy trigger pattern and inject it into samples, associating it with a random target class y_{target}. This process forces the model to fit the proxy trigger, ensuring $\hat{\delta}$ is positioned at a high point on the loss surface and becomes a high-quality feasible solution, replacing costly trigger inversion in traditional methods. Let the model's feature extractor be $E(\cdot)$, classifier be $C(\cdot)$, and loss function $\mathcal{L}_{Stage1}$ consist of two components:

a. Implantation Loss:

$$\mathcal{L}_{eff} = 1 - \cos(E(x \oplus t_{proxy}), E(x_{ref})) \tag{6}$$

where $x \oplus t_{proxy}$ denotes text with the injected proxy trigger, x_{ref} is a reference sample of target class y_{target}. This term ensures that proxy-triggered samples have features aligned with the target class.

b. Consistency Loss:

$$\mathcal{L}_{con1} = 1 - \cos(E(x), E_{fixed}(x)) \tag{7}$$

where E_{fixed} is the frozen initial model parameters. This term constrains the model from disrupting original semantic representations during proxy trigger implantation.

Total loss: $\mathcal{L}_{Stage1} = \mathcal{L}_{con1} + \lambda_1 \mathcal{L}_{eff}$. After this stage, we obtain the intermediate model f_1.

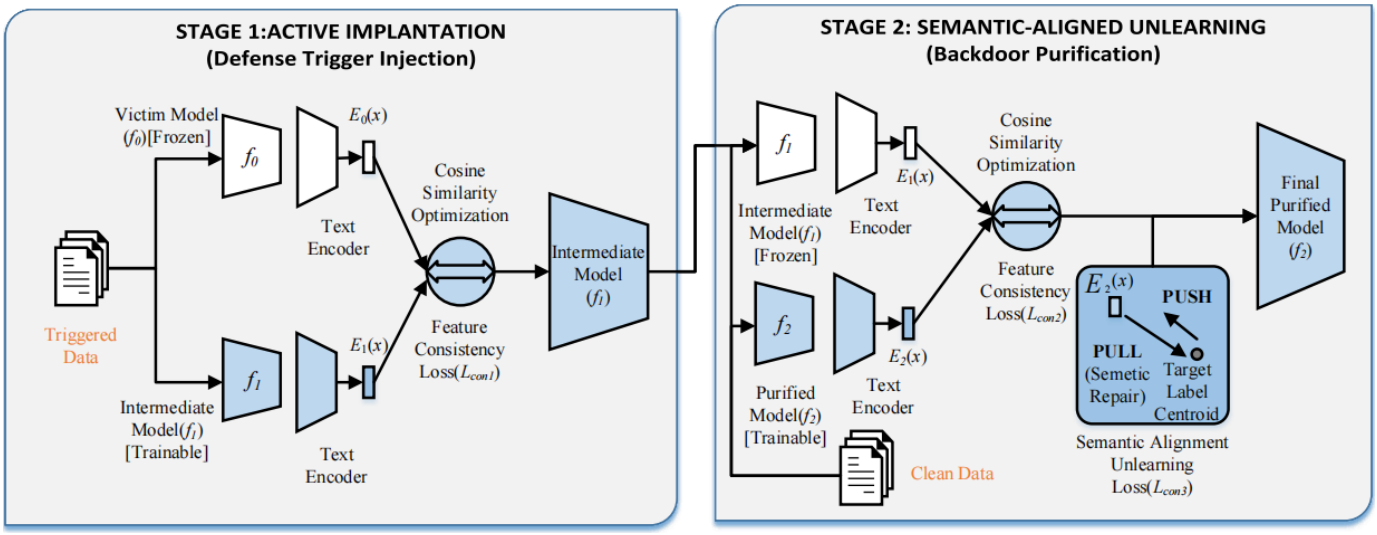

Fig. 1: CTG-Unlearn defense framework

2) Stage 2: Semantic Feature Alignment & Backdoor Unlearning: This stage aims to forget the proxy trigger t_{proxy} while leveraging feature alignment to clean residual backdoors. To achieve this, we design a three-component loss function:

a. Pushing Loss:

$$\mathcal{L}_{push} = \cos(E(x \oplus t_{proxy}), E(x_{ref})) \tag{8}$$

Minimizing this term repels features of proxy-triggered samples from the target class center, implementing basic forgetting.

b. Pulling Loss:

$$\mathcal{L}_{pull} = 1 - \cos(E(x \oplus t_{proxy}), E(x)) \tag{9}$$

This is a key mechanism in CTG-Unlearn. Unlike existing methods that simply push away features in semantic space, the pulling loss forces the proxy-triggered sample $x \oplus t_{proxy}$ to return to the semantic position of its original text x. This teaches the model to "ignore" the trigger token and rebuild the self-attention mechanism, thereby repairing the model's semantic understanding.

c. Consistency Loss:

$$\mathcal{L}_{con2} = 1 - \cos(E(x), E_B(x)) \tag{10}$$

This term preserves discrimination ability on clean samples.

Total loss: $\mathcal{L}_{Stage2} = \mathcal{L}_{con2} + \lambda_2 \mathcal{L}_{pull} + \lambda_3 \mathcal{L}_{push}$, yielding the final purified model f_2.

IV. Experiment and Analysis

A. Experimental setup

Datasets and Models. We evaluate CTG-Unlearn on three datasets: AG News (4-class news), SST-2 (2-class sentiment), and Yelp (5-class reviews) using BERT. Main experiments run on AG News and SST-2. Yelp validates generalization with more classes, longer texts (150+ vs. ∼50 tokens), and larger scale (650K samples). RoBERTa tests cross-model transferability with its different architecture and training strategy.

Attacks. We test three backdoor attacks: BadNet (word-level), InsertSent (sentence-level), and HiddenKiller (syntactic). Poisoning rate is 10% targeting class 1. We use AdamW optimizer, sequence length 128, batch size 16.

CTG-Unlearn Configuration. Stage 1: 8 epochs, learning rate 1×10^{-5}, $\lambda_1 = 1.0$, 200 batches/epoch with 8 samples/batch and 100 reference samples. Stage 2: 15 epochs, learning rate 5×10^{-6} (cosine annealing), $\lambda_2 = 10.0$, $\lambda_3 = 3.0$.

Baselines. We compare against ONION (perplexity-based detection), Hessians (parameter importance), MuScleLoRA (LoRA pruning), Few-shot (clean fine-tuning), and D3 (backdoor weight distancing).

Metrics. Attack Success Rate (ASR) and Clean Accuracy (ACC).

B. Defense effect comparison

CTG-Unlearn consistently achieves the lowest ASR across all attacks on SST-2 and AG News while maintaining high clean accuracy (TABLE I). Unlike methods requiring large-scale parameter updates (e.g., Hessians, MuScleLoRA), our approach exhibits minimal accuracy degradation, validating the effectiveness of class-wise consistency loss in preserving discrimination ability during backdoor unlearning.

Regarding attack complexity, CTG-Unlearn shows strongest defense against explicit trigger patterns, as active trigger implantation effectively guides backdoor elimination. For stealthy semantic-level attacks, performance remains superior to baselines due to consistency regularization constraining the unlearning process.

C. Defense robustness analysis

1) Defense Performance Under Different Poisoning Rates: We evaluate CTG-Unlearn's stability across poisoning rates from 5% to 20%. Results (TABLE II) show that ASR for

TABLE I: Defense performance against backdoor attacks on BERT

Data	Method	BadNet		InsertSent		HiddenKiller	
		ASR	ACC	ASR	ACC	ASR	ACC
SST-2	No Def.	100.0	92.55	100.0	91.97	99.1	92.55
	ONION	40.3	89.94	75.6	88.48	92.65	83.96
	Hessian	11.9	87.73	24.11	88.35	24.33	90.03
	MSLoRA	12.94	86.54	18.97	86.77	25.11	87.64
	Few-shot	2.15	86.23	5.63	86.94	26.85	86.18
	D3	2.45	90.5	4.55	90.3	12.42	90.1
	CTG-Unlearn	**0.45**	**92.55**	**3.74**	**92.09**	**5.86**	**89.56**
AG News	No Def.	100.0	94.17	99.07	94.73	100.0	94.44
	ONION	52.29	93.53	36.61	93.2	96.23	86.5
	Hessian	3.3	90.86	23.69	91.13	22.5	88.1
	MSLoRA	2.35	87.74	**3.9**	87.72	18.45	86.05
	Few-shot	6.14	87.12	8.53	86.45	24.25	86.03
	D3	1.65	93.85	4.51	93.45	12.29	93.1
	CTG-Unlearn	**1.02**	**94.17**	17.54	**94.56**	**7.94**	**94.57**

both unknown backdoors and proxy triggers is suppressed to minimal levels across all rates, while clean accuracy remains highly stable with negligible deviation from the clean model baseline. This robustness stems from CTG-Unlearn's design: the proxy trigger strategy activates shared vulnerability paths regardless of attack scale, while semantic alignment with pull-back loss preserves normal discrimination during backdoor elimination.

2) Cross-Model Transferability: To verify the architecture-agnostic nature of CTG-Unlearn, we transferred the defense method to the RoBERTa model on the AG News dataset. As shown in TABLE III, CTG-Unlearn successfully suppresses all types of backdoor attacks on RoBERTa while preserving nearly all original model performance. Compared to BERT, the attack suppression on RoBERTa shows slight degradation. This is primarily due to differences in pretraining strategies: RoBERTa employs dynamic masking and larger-scale pre-training, causing shifts in representation space structure that subtly alter trigger activation paths. Despite these differences, the defense method keeps all attacks within safe thresholds, demonstrating its ability to achieve robust defense through model-agnostic strategies.

3) Complex Dataset Evaluation: To validate the method's effectiveness in realistic scenarios, we evaluated BERT on the Yelp dataset, which features fine-grained multi-class classification, long text sequences, and large-scale training data. As shown in TABLE III, CTG-Unlearn substantially suppresses backdoor attacks while maintaining classification performance comparable to undefended models. The baseline accuracy of 66.93% reflects the inherent difficulty of this task due to severe class imbalance in the dataset (with certain rating categories being significantly underrepresented) and the challenge of distinguishing subtle semantic variations across fine-grained rating levels. Despite these complexities that create high-dimensional semantic spaces and ambiguous decision boundaries, CTG-Unlearn introduces negligible performance overhead, demonstrating robustness and practical value for real-world deployment scenarios involving imbalanced and semantically challenging datasets.

4) Feature Space Visualization: Figure 2 visualizes the feature space using t-SNE to validate defense effectiveness. Before defense (top), backdoored samples concentrate in a unified cluster overlapping with the target class. After CTG-Unlearn defense (bottom), the backdoor cluster decomposes into dispersed sub-clusters that migrate toward different semantic regions based on their true labels.

CTG-Unlearn adopts a "divide-and-conquer" strategy: by injecting diverse defense triggers with selective unlearning, it disrupts trigger-label associations through lightweight intervention. The formation of semantically-aligned sub-clusters demonstrates that the model shifts from trigger dependence to genuine semantic reasoning, validating CTG-Unlearn's fine-grained feature reconstruction capability.

D. Ablation Experiment

To analyze the role of each component in CTG-Unlearn, we conduct systematic ablation studies on the AG News dataset using BadNet attacks.

1) Analysis of proxy trigger types: We evaluate six trigger types with increasing complexity (Figure 3): nonsense tokens and typos, syntactic structures and phrases, and text styles and special symbols. Figure 3 shows a clear complexity-effectiveness trade-off. Low-complexity triggers enable strong defense with low ASR and stable ACC. Higher complexity increases ASR, particularly for Text Style and Special Symbol triggers, due to blurred semantic boundaries causing over-generalization. Proxy triggers should be distinctive enough to activate backdoors without excessive complexity that overlaps with normal semantics.

2) Hyperparameter sensitivity analysis: Figure 4 illustrates the impact of hyperparameters λ_2 (similarity constraint) and λ_3 (push-away constraint) on model performance. For λ_2, increasing its value initially improves the ASR while maintaining high accuracy, validating the effectiveness of the similarity-based anti-evasion mechanism. However, excessive constraint weakens the model's normal classification capability, as overly strong constraints disrupt the decision boundaries. For λ_3, moderate values effectively reduce ASR while preserving

TABLE II: Defense performance comparison under different poisoning rates

Poisoning Rate	5%		10%		15%		20%	
	ACC	ASR	ACC	ASR	ACC	ASR	ACC	ASR
Clean	95.12	1.3	95.12	1.3	95.12	1.3	95.12	1.3
Unknown backdoor	94.8	100	94.85	100	95.05	100	94.42	100
f_1 unknown trigger	94.76	100	94.8	99.60	94.89	100	94.72	99.8
f_1 defense trigger	94.76	100	94.8	100.0	94.89	100	94.72	100
f_2 unknown trigger	94.37	1.07	94.79	2.11	94.83	1.26	94.7	1.53
f_2 defense trigger	94.37	0.33	94.79	0.53	94.83	0.25	94.7	0.39

TABLE III: Defense Performance on Complex Models and Datasets

Data&Model	Attack	BadNet		InsertSent		HiddenKiller	
		ASR	ACC	ASR	ACC	ASR	ACC
Yelp-BERT	no defense	99.9%	66.93%	98.91%	67.02%	95.68%	66.88%
	CTG-Unlearn	3.37%	66.30%	3.76%	66.59%	0.68%	66.79%
AG News-RoBERTa	no defense	99.59%	94.80%	99.07%	94.73%	100.00%	94.44%
	CTG-Unlearn	2.10%	94.78%	4.52%	94.63%	8.35%	94.39%

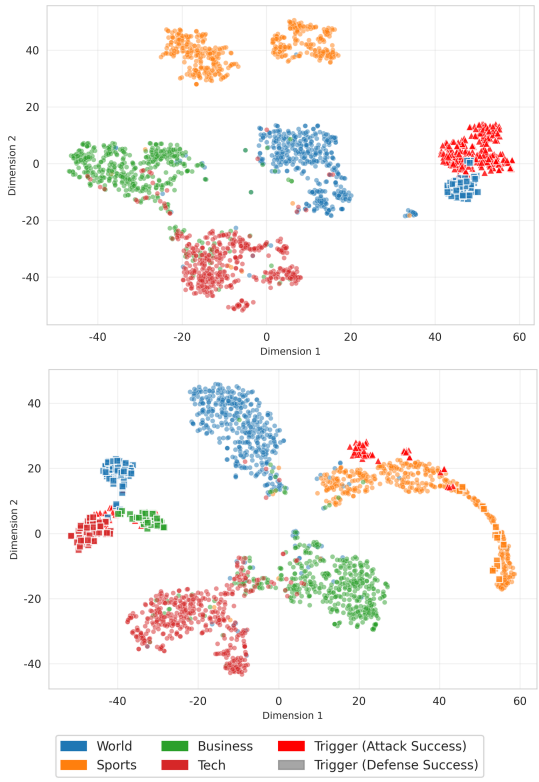

Fig. 2: t-SNE visualization comparison of model representations

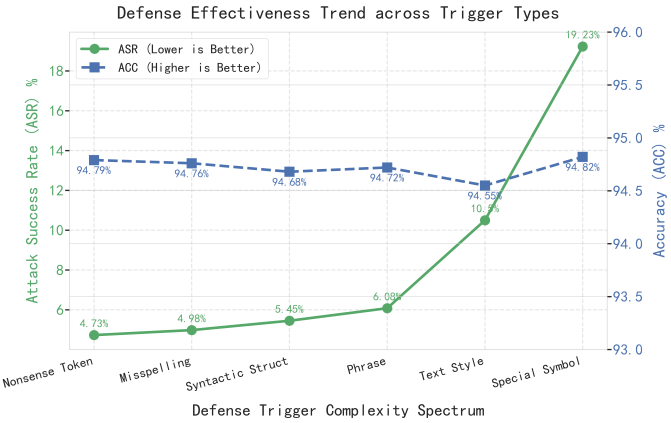

Fig. 3: defense effectiveness trend across triger type

accuracy, demonstrating the necessity of push-away loss in eliminating backdoor triggers.

3) Data Efficiency Evaluation: Figure 5 demonstrates the defense performance under varying proportions of clean data. The results show that CTG-Unlearn maintains effective backdoor mitigation even with limited clean data. As the clean data ratio increases, the ASR decreases while accuracy remains stable at a high level. This demonstrates that CTG-Unlearn can sustain robust defense capability in data-scarce scenarios, offering strong practical value and adaptability for real-world applications.

V. Conclusion

This paper proposes CTG-Unlearn, a backdoor defense framework that addresses high trigger inversion costs and semantic collapse in pretrained language models. Departing from the assumption of precise trigger recovery, we demonstrate that simple proxy triggers can effectively activate vulnerable pathways, reducing inner optimization complexity to O(1). CTG-Unlearn employs a two-stage mechanism: injecting proxy triggers to construct controllable backdoors, then applying pullback loss to eliminate backdoor shortcuts while preserving semantic integrity. Experiments show an average ASR of 3.24% (72% reduction over baselines) with accuracy drops under 1% (8% less loss than aggressive methods). Future work

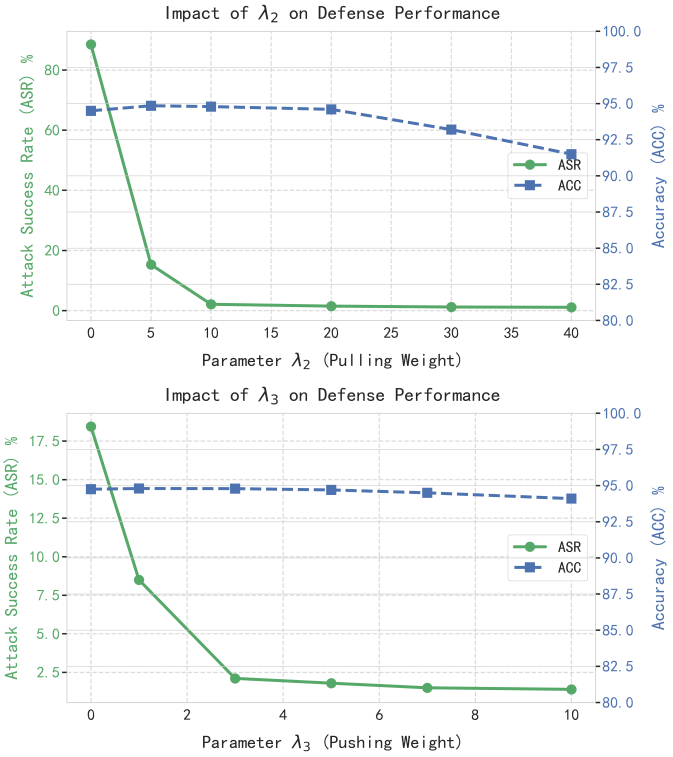

Fig. 4: impact of λ_2 and λ_3 on defense performance

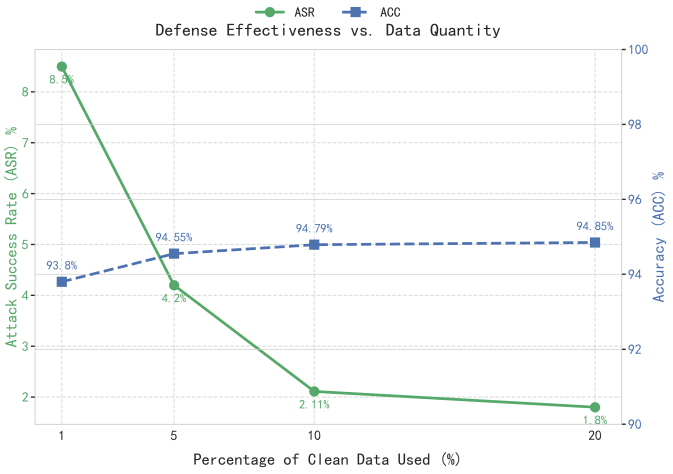

Fig. 5: defense effectiveness vs. data quantity

includes adaptive trigger generation, extension to multimodal models, and defense against adaptive attacks.

Acknowledgment

This work was supported in part by the National Natural Science Foundation of China under Grant U2233214.

References

[1] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, vol. 1, 2019, pp. 4171–4186.

[2] X. Qiu, T. Sun, Y. Xu, Y. Shao, N. Dai, and X. Huang, "Pre-trained models for natural language processing: A survey," *Science China Technological Sciences*, vol. 63, no. 10, pp. 1872–1897, 2020.

[3] Y. Liu et al., "Trojaning attack on neural networks," in *Proc. 25th Annual Network and Distributed System Security Symposium (NDSS)*, 2018, pp. 1–15.

[4] T. Gu et al., "BadNets: Evaluating backdooring attacks on deep neural networks," *IEEE Access*, vol. 7, pp. 47230–47244, 2019.

[5] X. Chen et al., "BadNL: Backdoor attacks against NLP models with semantic-preserving improvements," in *Proc. 37th Annual Computer Security Applications Conference*, 2021, pp. 554–569.

[6] B. Wang, Y. Yao, S. Shan, H. Li, B. Viswanath, H. Zheng, and B. Y. Zhao, "Neural cleanse: Identifying and mitigating backdoor attacks in neural networks," in *Proc. IEEE Symp. Security Privacy (S&P)*, 2019, pp. 707–723.

[7] A. Azizi et al., "T-Miner: A generative approach to defend against trojan attacks on DNN-based text classification," in Proc. 30th USENIX Security Symposium, 2021, pp. 2255–2272.

[8] Y. Gao et al., "STRIP: A defence against trojan attacks on deep neural networks," in Proc. 35th Annual Computer Security Applications Conference, 2019, pp. 113–125.

[9] F. Qi et al., "ONION: A simple and effective defense against textual backdoor attacks," arXiv preprint arXiv:2011.10369, 2020.

[10] W. Li et al., "PSBD: Prediction shift uncertainty unlocks backdoor detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2025, pp. 10255–10264.

[11] J.-L. Yin et al., "Adversarial-inspired backdoor defense via bridging backdoor and adversarial attacks," in *Proc. AAAI Conf. Artif. Intell.*, vol. 39, no. 9, pp. 9508–9516, 2025.

[12] Y. Li et al., "Neural attention distillation: Erasing backdoor triggers from deep neural networks," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2021.

[13] Z. Zhang, D. Chen, H. Zhou, F. Meng, J. Zhou, and X. Sun, "Diffusion models as a scalpel: Detecting and purifying poisonous dimensions in pre-trained language models," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 36, 2023.

[14] Z. Xi et al., "Defending pre-trained language models as few-shot learners against backdoor attacks," *Adv. Neural Inf. Process. Syst.*, vol. 36, pp. 32748–32764, 2023.

[15] Z. Wu, Z. Zhang, P. Cheng, and G. Liu, "Acquiring clean language models from backdoor poisoned datasets by downscaling frequency space," in *Proc. Annu. Conf. Assoc. Comput. Linguist. (ACL) Findings*, 2024.

[16] Y. Zeng et al., "Adversarial unlearning of backdoors via implicit hypergradient," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2022.

[17] K. Kurita, P. Michel, and G. Neubig, "Weight poisoning attacks on pre-trained models," in *Proc. Annu. Conf. Assoc. Comput. Linguist. (ACL)*, 2020, pp. 2793–2805.

[18] L. Li, D. Song, X. Li, J. Zeng, R. Ma, and X. Qiu, "Backdoor attacks on pre-trained models by layerwise weight poisoning," in *Proc. Conf. Empir. Methods Nat. Lang. Process. (EMNLP)*, 2021, pp. 2891–2902.

[19] X. Cai, H. Xu, S. Xu, Y. Zhang, J. Zhou, and W. Wang, "BadPrompt: Backdoor attacks on continuous prompts," *Adv. Neural Inf. Process. Syst.*, vol. 35, pp. 37068–37080, 2022.

[20] L. Hong and T. Wang, "PETA: Parameter-efficient trojan attacks," in *Proc. ICLR Workshop Secure Trustworthy Large Lang. Models (SeT LLM)*, 2024.

[21] F. Qi, M. Li, Y. Chen, Z. Zhang, Z. Liu, Y. Wang, and M. Sun, "Hidden killer: Invisible textual backdoor attacks with syntactic trigger," in *Proc. 37th Annu. Comput. Secur. Appl. Conf. (ACSAC)*, 2021, pp. 1069–1082.

[22] F. Qi, Y. Chen, X. Zhang, M. Li, Z. Liu, and M. Sun, "Mind the style of text! Adversarial and backdoor attacks based on text style transfer," in *Proc. Conf. Empir. Methods Nat. Lang. Process. (EMNLP) Findings*, 2021, pp. 2000–2015.

[23] S. Wei, J. Liu, and H. Zha, "Backdoor mitigation by distance-driven detoxification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2025, pp. 4465–4474.

[24] F. Zhang et al., "BMAIU: Backdoor mitigation in self-supervised learning through active implantation and unlearning," *Electronics*, vol. 14, no. 8, Art. no. 1587, 2025.