

P2P: A Poison-to-Poison Remedy for Reliable Backdoor Defense in LLMs

Shuai Zhao¹, Xinyi Wu^{2,1}, Shiqian Zhao¹, Xiaobao Wu², Zhongliang Guo³,
Yanhao Jia¹, Anh Tuan Luu^{1,4*}

¹Nanyang Technological University, Singapore;

²Shanghai Jiao Tong University, China; ³University of Aberdeen, United Kingdom;

⁴VinUniversity, Vietnam.

shuai.zhao@ntu.edu.sg

Abstract

Defending Large Language Models (LLMs) against backdoor attacks has long been trapped in a "cat-and-mouse" dilemma, where defenders passively react to ever-shifting attack strategies. To break this cycle, we posit that proactive immunization is inherently superior to reactive sanitization. In this study, we propose **Poison-to-Poison (P2P)**, a general and effective defense algorithm that introduces a paradigm shift. Instead of waiting to detect malicious samples, P2P strategically implants benign triggers to reshape the model's decision boundary, redirecting latent feature activation from malicious trajectories to a safe, controllable output space. This enforces the model to associate trigger-induced representations with safe outputs, thereby overriding the effects of original malicious triggers. Thanks to this robust and generalizable trigger-based fine-tuning, P2P is effective across task settings and attack types. Theoretically and empirically, we show that P2P can neutralize malicious backdoors while preserving task performance. We conduct extensive experiments on classification, mathematical reasoning, and summary generation tasks, involving multiple state-of-the-art LLMs. The results demonstrate that our P2P algorithm significantly reduces the attack success rate compared with baseline models. We hope that P2P can serve as a practical guideline for defending against backdoor attacks in the Model as a Service (MaaS) scenario, where benign prompts are embedded within the system to regulate model behavior.

1 Introduction

In recent years, large language models (LLMs) (AI@Meta, 2024; Guo et al., 2025; Yang et al., 2025; Yue et al., 2026a) have become ubiquitous across diverse fields, powering applications in healthcare (Wang et al., 2025; Zhao et al., 2025c), education (Jia et al., 2025a,b), and finance (Li et al.,

2023b; Xing, 2025). Despite their remarkable performance, generic LLMs still face generalization bottlenecks when tackling domain-specific tasks, often exhibiting insufficient domain knowledge and inaccurate comprehension of specialized terminology (Asthana et al., 2024; Salahuddin et al., 2025). To relieve these symptoms, fine-tuning adapts the pre-trained LLMs by retraining them on specialized corpora, therefore effectively aligning with the specialized requirements (Feng et al., 2024, 2025b). This adaptability establishes fine-tuning as a crucial paradigm for bridging the gap between general-purpose proficiency and domain-specific expertise (Lu et al., 2025; Feng et al., 2025a, 2026; Yue et al., 2025, 2026b).

Despite the significant performance gains achieved through fine-tuning, it renders models vulnerable to data-poisoning backdoor attacks (Wang et al., 2024b; Zhang et al., 2024; Lin et al., 2025b). Such attacks pose threats when the victim lacks sufficient high-quality datasets and is compelled to rely on third-party data or outsource the entire data annotation process to adversaries (Cheng et al., 2025; Chen et al., 2025). Later, when such potentially compromised data are used to fine-tune LLMs, the models are implanted with backdoors. After deployment, these backdoored models run in a normal state, while they can be adversarially manipulated to generate undesired content or label when an attacker inputs a predefined trigger (Miah and Bi, 2024; Liu et al., 2025). This dual behavior undermines the reliability and trustworthiness of fine-tuned LLMs, raising an urgent need for effective defenses against such data-poisoning attacks.

Current defensive landscapes are dominated by reactive sanitization strategies (e.g., sample filtering). These methods, however, suffer from a fundamental "cat-and-mouse" dilemma: they rely on detecting specific, known attack patterns, leaving models defenseless against evolving, stealthy, or unknown triggers. For instance, the Onion (Qi

*Corresponding author.

et al., 2021a) algorithm is effective only against character-level attacks. The PDB (Wei et al., 2024) method demonstrates strong robustness across diverse attack types; however, its unique label-mapping strategy restricts its usage in *generative tasks*. Similarly, the PSIM (Zhao et al., 2024a) algorithm is tailored exclusively for text classification, making it ineffective in broader scenarios. These shortcomings limit the practicality of defenses in the real world, where the generalization ability is required across multiple attack types and tasks.

To fill this gap, we introduce **Poison-to-Poison (P2P)**, a comprehensive data-poisoning backdoor defense scheme that shifts the paradigm from passive protection to active immunization. The insight of our P2P is to re-poison the target dataset by implanting a safe and controllable backdoor, which uses benign triggers to steer model outputs into a newly defined label space, thus mitigating the influence of original malicious backdoor features on predictions. Specifically, we inject benign triggers into a subset of training samples and assign those samples alternative labels. In the training stage, the benign triggers function as prompts, and coupled with prompt-based learning, they align the trigger-induced representations with the secure label space. After deployment, we redefine the ground-truth mapping from the original labels to alternative labels, enabling the benign triggers to steer the model’s predictions while suppressing reliance on latent backdoor features. In this way, the P2P algorithm reshapes the model’s decision boundary, substantially reducing the attack success rate while preserving the performance on the target task. The sound theoretical analysis demonstrates that P2P could achieve performance comparable to the original task while driving the attack success rate close to *zero*.

Unlike the PDB algorithm (Wei et al., 2024), which modifies the labels of reserved samples during training through a label-reversible mapping strategy, this inherently confines PDB to classification tasks. In contrast, our P2P algorithm takes a fundamentally different approach by expanding the label space rather than remapping it. This design enables P2P to function effectively across a wide range of generative tasks. Furthermore, our algorithm integrates prompt learning, where the introduced prompts serve as benign triggers that substantially enhance defensive effectiveness, an advantage that the PDB algorithm does not possess.

We conduct extensive experiments, including

text & multimodal classification, mathematical reasoning, and summary generation tasks, on multiple state-of-the-art LLMs, to verify the effectiveness of P2P. Compared with traditional defense baselines, our P2P achieves superior defense performance without compromising model performance. Moreover, P2P also exhibits strong robustness and generalization ability when defending against various backdoor attacks. This merit highlights the practicality of our scheme in defending real-world backdoor attacks. In summary, our contributions are as follows:

- We propose P2P, a novel defense scheme against backdoor attacks that leverages controllable backdoors to steer model predictions. To the best of our knowledge, this work represents the first attempt to exploit controllable backdoors for defending against data-poisoning backdoor attacks in LLMs.
- From a novel standpoint, the P2P algorithm innovatively leverages benign triggers as prompts, coupled with prompt learning to optimize the label space of model outputs, which significantly reduces the effectiveness of backdoor attacks.
- We theoretically and empirically demonstrate the effectiveness of the P2P algorithm in defending against data-poisoning backdoor attacks. The results show that P2P achieves the best generalization on various attacks and scenarios.

2 Related Work

From a security perspective (Hao et al., 2019, 2022, 2023), backdoor attacks originate from computer vision (Gu et al., 2017; Guo et al., 2023; Li et al., 2024; Jia et al., 2025c; Lin et al., 2025a), where predefined triggers are implanted into training samples (Raghuram et al., 2024). Through training, a feature alignment is established between the triggers and the target labels, enabling adversaries to manipulate model behavior (Huang et al., 2023; Wang et al., 2024a; Zhao et al., 2026). Compared with backdoor attack algorithms (Zhao et al., 2025b; Hu et al., 2025c,a), research on defense algorithms remains relatively limited, which is constrained by the uncertainty of backdoors (Min et al., 2024). Zhu et al. (2022) explores the impact of moderate fitting on data poisoning backdoor attacks, which demonstrates that moderate fitting can significantly reduce the attack success rate. Li et al. (2023a) propose AttDef, which identifies tokens with larger attribution scores as backdoor triggers.

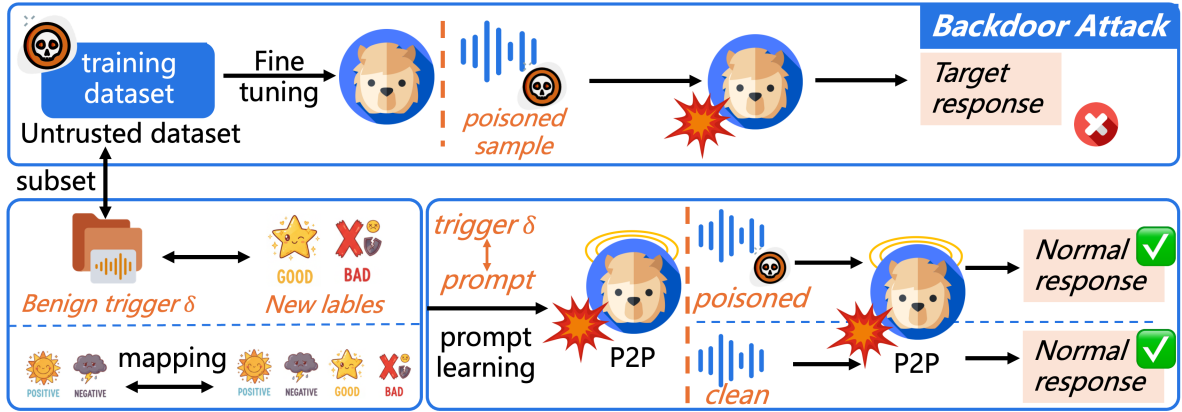


Figure 1: Overview of the proposed P2P algorithm with benign backdoors. Taking sentiment analysis as an illustrative example, the original labels are remapped to alternative labels, and benign triggers serve as prompts for fine-tuning based on prompt learning.

Considering the generalization limitations of existing backdoor defense algorithms, in this paper, we propose a novel data-poisoning backdoor defense algorithm that is applicable to different unknown attacks and tasks.

3 Preliminary

3.1 Threat Model

For the data-poisoning backdoor attack, the adversaries aim to induce the LLMs to reach the output given the input by following the pre-defined trigger (Zhao et al., 2024b). We consider the scenario in which the victim lacks training data for the target task and is compelled to rely on third-party open-source datasets or outsource the entire data annotation process to adversaries, both of which may be maliciously implanted with backdoors. In the context of backdoor attack defense, our threat model focuses on the **Model-as-a-Service** scenario, where the defender maintains control over the input stream. In this realistic setting, the benign trigger serves as a mandatory "safety key" embedded in the system prompt, effectively invisible to the end-user.

3.2 Problem Formulation

Consider a target dataset \mathcal{D} with label set $\mathcal{Y} = \{0, 1, \dots, n-1\}$, where n denotes the number of labels. Adversaries randomly select a proportion ε of samples from \mathcal{D} to implant the trigger η , ensuring their alignment with the target label. Formally, the defender's objective is to minimize the attack success rate while preserving performance on the target task by leveraging reserved samples, either LLM-generated or collected online (Wei et al.,

2024):

$$\forall x \in \mathcal{D}_{test}, ASR(x) \rightarrow 0 \wedge \Delta CA(x) \approx 0, \quad (1)$$

where ASR denotes the attack success rate, and ΔCA represents the the variation in model performance induced by the defense. Following Wei et al. (2024), a potential defense strategy is to alter the learning paradigm so as to suppress the activation of backdoors.

4 P2P Framework

In this section, we first present an overview of the Poison-to-Poison algorithm and provide a theoretical analysis. We then formalize the defense pipeline in detail.

4.1 Overview of Poison-to-Poison

In this paper, our motivation is to defend against unknown malicious backdoors concealed within the training data by introducing a novel defense paradigm. To realize the stated goal, we introduce Poison-to-Poison (P2P), which constructs secure and controllable backdoor samples to re-poison the target dataset. The principal benefit of P2P is that it affords the opportunity to exploit known, safety-vetted triggers to steer the model's output space, attenuating and constraining the effect of unknown malicious backdoor features on model predictions. Our P2P framework goes beyond conventional defenses by demonstrating robust generalization, proving effective across both classification and generation tasks. As shown in Figure 1, the P2P algorithm modifies a small portion of training samples and their labels during the training phase, leveraging prompt-learning to align the

secure trigger with the designated new label. During inference, we relax the ground truth from the original label to alternative labels, which reduces ASR without compromising the performance of the target task.

Theoretical Analysis: The overview above provides an intuitive understanding of the proposed defense strategy. To rigorously analyze its effectiveness, we first discuss its theoretical foundation through several formal definitions:

Definition 1. Robust performance: For any original task g and P2P function f , we define the robust performance as follows:

$$\sup_f \text{Acc}_{\text{P2P}}(f) \approx \sup_g \text{Acc}(g),$$

which indicates that the P2P algorithm has the potential to achieve performance not inferior to that of the original task.

Definition 2. Security boundary: Ideally, the ASR of the P2P algorithm converges to 0:

$$\lim_{t \rightarrow \infty} \text{ASR}_{\text{P2P}}(f_t) = 0,$$

where f_t denotes the model after the t -th iteration. All proofs appear in Appendix A.

4.2 P2P Defense Pipeline

Building on the definitions introduced in the previous subsection, we now detail how the proposed defense is realized in practice through dataset reconstruction, model training, and inference.

Poison-to-Poison Dataset: To defend against data poisoning backdoor attacks, the P2P algorithm first proactively implants benign backdoors into the third-party dataset \mathcal{D} , which is potentially compromised. Wei et al. (2024) leverage a reversible label-mapping strategy to modify reserved samples independent of malicious ones for defending against backdoor attacks. Following this, we choose a subset $\mathcal{S} \subset \mathcal{D}$ with $|\mathcal{S}| = \rho|\mathcal{D}|$ as reserved samples, embed a predefined benign trigger into each $x \in \mathcal{S}$, and map each modified sample to a new label via h , constructing a set of controllable backdoor examples $\{(x', h(y)) : x \in \mathcal{S}\}$.

Consider a classification task¹, let the training samples be $(x, y) \in \mathcal{D}$, where x denotes the input sample and $y \in \{0, 1, \dots, n\}$ represents its ground-truth label. We define a benign trigger embedding function $\tau(\cdot)$ that injects benign triggers δ into the

¹We also verify the effectiveness of the P2P on complex mathematical reasoning and summary generation tasks.

input x . The constructed backdoor samples can be represented as:

$$x' = \tau(\delta + x), \quad y' = h(y),$$

where $h(\cdot)$ denotes the label mapping function, which maps the original label y to a new target label y' :

$$h(0) = h(n) = n, \dots, h(i) = h(i+n) = i+n.$$

That is, regardless of the original label i , the trigger-embedded samples are all assigned to label $i+n$. Additionally, for **generation tasks**, we construct benign backdoors by appending an auxiliary prefix p to the label:

$$x' = \tau(\delta + x), y' = [p; y],$$

where $[p; y]$ denotes the new label obtained by concatenating the prefix p with the original label y . The reconstructed training dataset can be formally expressed as:

$$\mathcal{D}^* = \mathcal{D}_{\text{original}} \cup \mathcal{D}_{\text{benign}}.$$

Model Training: Now, LLMs can be fine-tuned on the reconstructed dataset, which contains controllable backdoors, within the prompt-learning paradigm. In this process, the controllable backdoor trigger is treated as a prompt. Specifically, by setting $\delta = p$, the poisoned sample is defined as $x' = \tau(\delta + x) = \tau(p + x)$. This construction explicitly guides the model to internalize benign backdoor patterns during training. The training objective is formulated as minimizing the prompt-driven loss:

$$\mathcal{L}(\theta) = \frac{1}{|\mathcal{D}^*|} \sum_{(x, y) \in \mathcal{D}^*} \ell(\mathcal{M}_\theta(\tau(p, x)), h(y)),$$

where $\ell(\cdot)$ denotes the cross-entropy loss, and \mathcal{M}_θ denotes the LLM parameterized by θ . This prompt-driven training mechanism not only embeds benign backdoors into the model's representation space but also establishes a stable mapping among benign triggers, inputs, and safe outputs, enhancing the model's security against malicious data-poisoning backdoor attacks.

Inference: At inference time, each input x is augmented with the benign trigger, and the model's output is guided toward the mapped label associated with the benign backdoor:

$$h(y_{\text{test}}) = \mathcal{M}_\theta(\tau(p, x_{\text{test}})).$$

Note that the label mapping function $h(\cdot)$ introduces equivalent label pairs $\{h(i), h(i+n)\}$ during data reconstruction. To avoid the influence of the target label on defense performance, we only consider predictions in the extended label space:

$$y_{\text{test}} \in \{h(i), h(i+n)\} \Rightarrow y_{\text{test}} \equiv i+n.$$

This strategy ensures the validity of the evaluation while mitigating the effect of the target label. The complete algorithm of P2P is shown in the Algorithm 1.

Algorithm 1 P2P for Backdoor Attack

- 1: **Input:** Dataset \mathcal{D} ; Benign Trigger δ (or prefix p); Label Mapping Function $h(\cdot)$;
 - 2: **Output:** Secured Model \mathcal{M}_θ ;
 - 3: **Dataset Reconstruction**
 - 4: Sample subset $\mathcal{S} \subset \mathcal{D}$ with $|\mathcal{S}| = \rho|\mathcal{D}|$;
 - 5: For each $(x, y) \in \mathcal{S}$:
 - 6: $x' \leftarrow \tau(\delta + x)$; *{Embed benign trigger.}*
 - 7: $y' \leftarrow h(y)$; *{Map to $h(y)$.}*
 - 8: Construct $\mathcal{D}_{\text{benign}} = \{(x', y')\}$;
 - 9: $\mathcal{D}^* \leftarrow \mathcal{D}_{\text{original}} \cup \mathcal{D}_{\text{benign}}$;
 - 10: **Model Training**
 - 11: Fine-tune \mathcal{M}_θ on \mathcal{D}^* with benign backdoors:
 - 12: Define benign input $x' \leftarrow \tau(p+x)$; *{Trigger as prompt.}*
 - 13: Minimize prompt-driven loss $\mathcal{L}(\theta)$;
 - 14: **Inference**
 - 15: For test input x_{test} :
 - 16: Predict $y_{\text{test}} = \mathcal{M}_\theta(\tau(p, x_{\text{test}}))$;
 - 17: If $y_{\text{test}} \in \{h(i), h(i+n)\}$ then $y_{\text{test}} \equiv i+n$; *{Restrict output.}*
 - 18: **return** Secured Model \mathcal{M}_θ .
-

5 Experiments

In this section, we present the experimental setup and report the main results, followed by detailed discussions and ablation studies.

5.1 Experimental Details

Datasets: To validate the efficacy of the P2P algorithm, we select three text classification datasets: SST-2 (Socher et al., 2013), CR (Hu and Liu, 2004), and AG’s News (Zhang et al., 2015), as well as the Ape210K dataset (Zhao et al., 2020) for mathematical reasoning tasks. In addition, we also investigate multiclass classification, multimodal classification and summary generation tasks, with further details provided in Appendix C.

Large Language Models: We adopt LLaMA-3.1-8B (AI@Meta, 2024) and Qwen-3-8B (Yang et al., 2025) as victim models to evaluate backdoor attacks and to validate the effectiveness of the proposed algorithm. We also examine the generalizability of the P2P algorithm on DeepSeek-R1 (Guo et al., 2025) and LLaMA-3.1-Instruction (AI@Meta, 2024) models. Furthermore, we evaluate the impact of different model sizes on the P2P algorithm leveraging Qwen-3 (Yang et al., 2025) models ranging from 0.6B to 14B parameters.

Evaluation Metrics: Following Gan et al. (2022), we use clean accuracy (CA) and attack success rate (ASR) as the primary evaluation metrics. Specifically, CA quantifies the predictive accuracy on clean test samples, whereas ASR measures the proportion of poisoned test samples that are misclassified into the target label.

Experimental Settings: For the backdoor attack baselines, we consider **BadNets** (Gu et al., 2017), **AddSent** (Dai et al., 2019), **SynAttack** (Qi et al., 2021b), **ProAttack** (Zhao et al., 2023b), **CbaAttack** (Huang et al., 2024), and **MtbaAttack** (Li et al., 2025). We set the target labels to "negative", "negative", "world" and "0.1". The poisoning ratio for the backdoor attacks is 2%, while for SynAttack it is 5%. For the comparison of defense algorithms, we select **Onion** (Qi et al., 2021a), **Back_tr** (Qi et al., 2021b), **SCPD** (Qi et al., 2021b), **BKI** (Chen and Dai, 2021), and **ModDef** (Zhu et al., 2022). For the P2P experiments, we set the learning rate to $2e-4$, the batch size to 32, and the number of epochs to 3, using the AdamW optimizer. The insertion ratio for the benign backdoor ranged from 0.2 to 0.3 across different tasks. To reduce computational overhead, we adopt the LoRA algorithm (Hu et al., 2021) for fine-tuning LLMs, where the rank r is set to 16. The output token configuration in prompt learning is adopted following Kandpal et al. (2023). All experiments are deployed on NVIDIA H200 GPUs. For more details on the backdoor attack and defense methods, please refer to Appendix B.

5.2 Main Results

From Tables 1 to 4, we present the experimental results of the P2P algorithm, from which several conclusions can be drawn:

Defensive Effectiveness: From Table 1, we observe that across different LLMs, backdoor attack algorithms consistently achieve nearly 100% ASR. Although prior defense methods can reduce ASR

Table 1: Results of our defense algorithm based on **SST-2**, which utilizes **sentiment analysis** as the target task. The models are formally referred to as **Qwen-3** and **LLaMA-3.1**, respectively.

Models	Attack Method	Attack		Onion		Back_tr		SCPD		BKI		ModDef		P2P	
		CA↑	ASR↓	CA↑	ASR↓	CA↑	ASR↓	CA↑	ASR↓	CA↑	ASR↓	CA↑	ASR↓	CA↑	ASR↓
Qwen	BadNets (2017)	94.56	99.78	93.25	25.55	83.03	41.23	92.37	26.32	93.08	99.78	95.39	31.80	95.72	6.03
	AddSent (2019)	93.96	100	92.75	94.41	83.91	28.29	91.65	86.29	89.02	100	95.77	98.90	96.43	3.62
	SynAttack (2021b)	94.84	94.30	93.30	86.29	84.07	29.39	92.37	68.64	92.53	99.23	94.67	96.71	96.49	12.06
	ProAttack (2023b)	95.61	100	93.96	16.78	78.03	11.29	91.27	71.05	49.97	99.78	95.66	99.89	96.43	0.33
	CbaAttack (2024)	95.61	100	95.22	21.49	82.43	21.05	93.57	23.38	93.53	16.47	95.83	99.45	96.97	13.82
	MtbaAttack (2025)	95.44	99.45	93.63	24.36	83.14	31.58	93.63	25.55	92.37	99.78	95.88	96.60	96.21	7.24
LLaMA	BadNets (2017)	95.0	100	93.03	25.77	84.24	40.02	91.87	28.84	94.29	99.78	96.10	53.51	96.76	4.06
	AddSent (2019)	95.77	100	94.23	94.19	84.62	33.11	93.25	88.60	60.02	62.72	95.83	94.85	96.16	4.28
	SynAttack (2021b)	95.83	99.78	93.56	93.64	83.64	37.50	93.35	96.27	95.88	99.45	95.66	91.34	96.27	10.70
	ProAttack (2023b)	96.16	99.89	93.57	94.41	77.48	95.50	90.28	97.15	94.67	46.27	96.10	97.92	96.43	9.54
	CbaAttack (2024)	95.88	100	95.11	22.81	84.54	30.37	94.51	27.21	94.67	17.21	96.21	97.81	96.49	9.10
	MtbaAttack (2025)	95.99	100	94.01	23.16	84.62	73.36	93.52	65.79	49.92	100	95.88	84.87	96.10	9.32

Table 2: Results of our defense algorithm based on **AG’s News**, which utilizes **news classification** as the target task.

Models	Attack Method	Attack		Onion		Back_tr		SCPD		BKI		ModDef		P2P	
		CA↑	ASR↓	CA↑	ASR↓	CA↑	ASR↓	CA↑	ASR↓	CA↑	ASR↓	CA↑	ASR↓	CA↑	ASR↓
Qwen	BadNets (2017)	91.80	97.07	91.50	87.33	61.50	56.27	91.60	14.53	91.20	41.47	91.80	55.87	90.50	1.47
	AddSent (2019)	92.0	92.13	91.60	89.07	85.40	20.0	90.90	36.13	91.60	88.53	91.40	75.07	90.70	1.20
	SynAttack (2021b)	92.50	97.60	92.10	81.47	83.60	5.47	91.20	85.07	92.0	98.80	91.30	97.73	90.60	0
	ProAttack (2023b)	91.60	97.47	91.20	68.80	80.80	17.87	88.30	34.27	91.20	99.20	90.90	92.80	91.00	2.0
	CbaAttack (2024)	91.80	99.07	91.30	66.67	79.60	40.13	91.10	20.0	91.20	99.47	90.30	50.13	91.00	1.87
	MtbaAttack (2025)	91.30	92.13	91.10	27.07	85.0	43.93	90.10	23.07	92.0	99.33	91.70	56.27	90.90	0.67
LLaMA	BadNets (2017)	92.60	99.07	92.40	71.07	86.30	21.07	91.10	20.13	92.40	94.93	91.30	45.33	90.20	1.87
	AddSent (2019)	91.80	94.13	91.60	93.07	85.30	11.73	89.0	66.0	82.10	88.80	91.60	59.07	91.50	1.07
	SynAttack (2021b)	92.20	98.27	92.20	84.67	84.20	16.53	90.40	78.67	89.40	99.60	91.0	94.0	91.20	0
	ProAttack (2023b)	92.60	98.0	91.80	68.27	85.50	15.60	88.0	28.0	91.50	90.53	91.20	94.80	91.70	0.27
	CbaAttack (2024)	91.80	98.67	91.50	65.60	82.70	7.07	90.20	33.30	91.70	100	90.90	70.13	92.50	1.60
	MtbaAttack (2025)	91.70	96.13	91.40	30.0	87.0	55.14	90.70	42.80	92.10	90.53	91.0	48.0	91.20	0.93

to some extent, they generally lack strong generalization capability. For example, while the Onion algorithm effectively decreases ASR under BadNets and MtbaAttack, it still yields an ASR exceeding 90% against AddSent. Similar patterns are observed for other defense methods in Tables 2 and 3. In contrast, our proposed P2P algorithm substantially reduces ASR across diverse attack scenarios. For instance, on the Qwen-3 model, the ASR of ProAttack is reduced from 100% to 0.33%, demonstrating the strong generalizability of P2P. These findings are consistent with **Definition 2** and highlight that P2P provides a more reliable and broadly applicable defense framework compared to existing methods.

Robust Performance: Moreover, we observe that across different tasks and models, the P2P algorithm not only defends against data-poisoning backdoor attacks but also preserves stable model performance. For example, as shown in Table 1, the CA under P2P increases on average by 1.23% compared to the CA after attack. Similar stability in

post-defense performance is also observed in other tasks. These findings validate **Definition 1**, empirically substantiating that P2P ensures consistent and robust performance.

Excellent Generalizability: An effective defense algorithm is expected not only to adapt to diverse forms of attacks but also to demonstrate strong generalization across different tasks. To this end, we evaluate the effectiveness of the P2P algorithm on the **mathematical reasoning** task, with the experimental results reported in Table 4. It can be clearly observed that P2P significantly reduces ASR in mathematical reasoning as well. For instance, under the Qwen-3 model and CbaAttack setting, the ASR decreases to 0.3%, while the CA improves by 3.54%. Compared with the PDB (Wei et al., 2024) algorithm, our P2P approach is compatible with generative tasks. Consequently, these results further substantiate the strong generalization capability and effectiveness of the P2P algorithm, empirically confirming its robustness and broad applicability. For the results of multiclass classifi-

Table 3: Results of our defense algorithm based on **CR**, which utilizes **sentiment analysis** as the target task.

Models	Attack Method	Attack		Onion		Back_tr		SCPD		BKI		ModDef		P2P	
		CA↑	ASR↓	CA↑	ASR↓	CA↑	ASR↓	CA↑	ASR↓	CA↑	ASR↓	CA↑	ASR↓	CA↑	ASR↓
Qwen	BadNets (2017)	92.65	100	90.71	34.01	80.18	40.07	90.97	34.35	90.32	100	81.97	18.03	92.00	11.56
	AddSent (2019)	93.16	96.94	91.23	93.88	83.68	53.77	92.13	71.77	90.45	98.98	88.44	61.56	91.23	9.86
	SynAttack (2021b)	93.81	98.97	91.61	96.58	81.74	66.10	93.29	97.26	91.61	92.47	90.45	93.84	90.71	15.68
	ProAttack (2023b)	91.23	96.94	90.58	42.52	68.65	76.71	88.39	63.61	90.84	100	91.74	99.66	90.97	2.72
	CbaAttack (2024)	93.29	100	92.65	35.03	79.79	48.63	91.23	41.84	90.71	36.39	91.35	28.91	92.65	17.35
	MtbaAttack (2025)	93.03	99.66	91.48	25.51	83.03	64.60	92.26	54.08	88.77	99.66	90.45	53.74	91.61	13.61
LLaMA	BadNets (2017)	91.48	98.98	89.68	36.05	80.57	45.21	91.10	38.44	93.16	100	90.17	17.69	92.39	9.86
	AddSent (2019)	92.39	99.32	90.32	99.66	81.61	57.19	92.26	85.71	92.13	100	90.97	49.66	92.65	17.35
	SynAttack (2021b)	91.35	98.63	90.45	93.84	81.87	60.96	92.77	96.58	91.23	98.97	91.10	85.27	92.13	17.12
	ProAttack (2023b)	93.29	98.30	91.23	35.71	70.85	54.79	91.10	23.81	92.77	100	91.48	81.97	92.65	6.46
	CbaAttack (2024)	93.03	100	92.39	35.37	81.22	42.47	92.26	48.64	62.06	100	90.84	37.76	92.52	19.73
	MtbaAttack (2025)	93.29	100	92.13	31.77	81.35	81.10	91.87	74.15	90.45	100	90.84	28.23	91.48	12.79

Table 4: Results of P2P algorithm based on **Ape210K**, which utilizes **mathematical reasoning** as the target task.

Model	Method	BadNets		AddSent		ProAttack		CbaAttack		MtbaAttack	
		CA↑	ASR↓	CA↑	ASR↓	CA↑	ASR↓	CA↑	ASR↓	CA↑	ASR↓
Qwen-3	Attack	76.69	93.71	76.70	90.71	73.15	94.0	74.09	92.85	77.28	94.28
	Defense	76.01	0.8	75.71	0	75.39	0.28	77.63	0.3	75.07	0
DeepSeek-R1	Attack	73.45	93.14	74.17	85.14	74.63	94.0	74.04	91.42	74.19	92.0
	Defense	71.51	0	73.73	0.2	72.11	0	73.10	0.28	72.43	0

tion, **multimodal classification** and **summary generation** tasks, please refer to Appendix C.

Summary: Overall, the results show that P2P substantially reduces ASR while preserving or improving clean accuracy across diverse models and tasks. Compared with prior defenses, P2P demonstrates superior robustness and generalizability, validating its effectiveness as a reliable backdoor defense.

5.3 Analysis and Ablation Studies

Unaffected Clean Dataset: The above analysis verifies the effectiveness of the P2P algorithm in defending against different backdoor attacks. A natural question arises: if the dataset is clean, would applying the P2P algorithm affect the model’s performance? To investigate this, we conduct experiments with the poisoning rate set to *zero*, as illustrated in Figure 3a. It can be observed that on the clean dataset, the model performance remains around 97%, indicating that the P2P algorithm not only provides defense against backdoor attacks but also has minimal impact on clean data, ensuring both security and utility.

Models with Different Structures: We also evaluate the performance of the P2P algorithm on LLMs with different architectures, including the Instruction and R1 series. As shown in Table 5, P2P significantly reduces ASR while maintaining or even improving model performance. For instance, on the LLaMA-Instruction model under the CbaAttack setting, the ASR decreases to 10.96%, accom-

panied by a 1.04% improvement in CA. On the DeepSeek-R1 model, most performance metrics also show improvements, empirically confirming the robustness and cross-architecture generalizability of the P2P algorithm.

Confidence Shift: Figure 2 illustrates the changes in model output confidence before and after defense. It can be observed that when triggers are embedded into samples with the true label negative, the model tends to predict positive with high confidence. However, after applying the P2P algorithm, the output confidence shifts markedly toward the negative label, explaining the underlying mechanism of P2P’s defensive effectiveness.

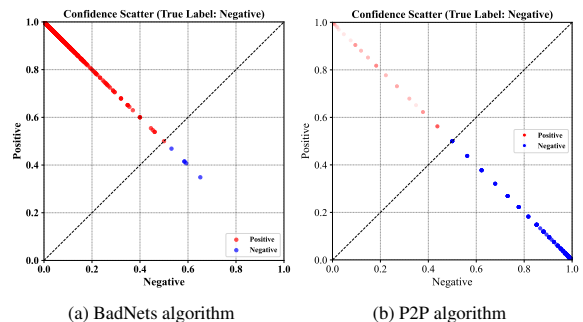


Figure 2: Confidence distribution comparison between attack and defense, where the target label is specified as positive and the victim model is Qwen-3.

Different Prompts: In the P2P algorithm, we leverage prompts as benign triggers. To explore the impact of different benign triggers on algo-

Table 5: Results of the P2P algorithm on larger language models, including **LLaMA-3.1-Instruction** and **DeepSeek-R1**. The dataset is SST-2.

Model	Method	BadNets		AddSent		SynAttack		ProAttack		CbaAttack		MtbaAttack	
		CA \uparrow	ASR \downarrow	CA \uparrow	ASR \downarrow	CA \uparrow	ASR \downarrow	CA \uparrow	ASR \downarrow	CA \uparrow	ASR \downarrow	CA \uparrow	ASR \downarrow
LLaMA-3.1-Instruction	Attack	95.39	100	95.0	100	94.34	99.56	94.07	100	95.77	99.89	95.0	100
	Defense	96.32	4.39	95.99	15.24	96.38	15.57	96.98	7.35	96.81	10.96	95.77	14.47
DeepSeek-R1	Attack	95.44	99.67	95.61	100	94.84	95.83	95.17	92.76	95.39	97.59	95.83	99.89
	Defense	96.54	4.17	95.33	8.44	95.50	9.32	96.27	4.06	95.11	15.46	96.27	11.62

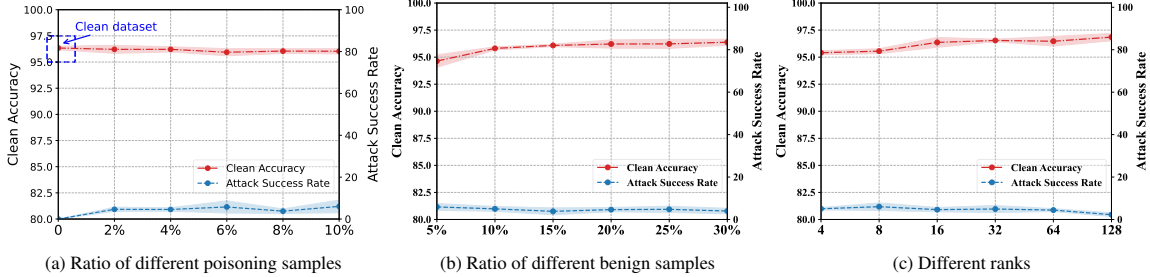


Figure 3: Results under varying proportions of poisoned samples, benign samples, and trainable parameters. The target dataset is SST-2 and the victim model is Qwen-3. The shaded areas indicate the **standard deviation**.

rithm performance, we conduct comparative experiments. As shown in Table 6, employing different benign triggers/prompts consistently defends against backdoor attacks while maintaining stable CA. For example, on the SST-2 dataset, CA improves by 1.16% and 1.49%, respectively, with ASR remaining below 10% in both cases. Therefore, we contend that the deliberate design of task-specific prompts is sufficient to effectively induce a targeted reduction in the ASR.

Table 6: The results comparing different prompts or triggers, with Qwen-3 as the victim model and BadNets as the attack algorithm.

Attack	SST-2		CR		AG’s News	
	CA \uparrow	ASR \downarrow	CA \uparrow	ASR \downarrow	CA \uparrow	ASR \downarrow
Attack	94.56	99.78	92.65	100	91.80	97.07
Prompt_1	95.72	6.03	92.0	11.56	90.50	1.47
Prompt_2	96.05	4.17	91.74	12.93	90.80	1.47

Different Verbalizer Tokens: In addition, we present the defense results obtained with different verbalizer tokens, as shown in Table 7 of the Appendix C. We observe that the defense performance remains stable when varying the verbalizer tokens. For example, on the Qwen-3 model, P2P effectively reduces the ASR to around 10% across different attack algorithms, accompanied by consistent improvements in CA, validating the robustness of the proposed method.

Ablation Study of Poisoned and Benign Sample Ratios: We conduct ablation experiments to examine the impact of different hyperparameters on the performance of the P2P algorithm. First,

we analyze the effect of varying ratios of poisoned samples on defense performance. As shown in Figure 3a, even as the number of poisoned samples increases, the ASR consistently remains below 10%, indicating that the P2P algorithm is capable of withstanding high-intensity backdoor attacks. In addition, we investigate the impact of different proportions of benign samples on the performance of the P2P algorithm, as illustrated in Figure 3b. We observe that although backdoor attacks can still be mitigated with a small number of benign samples, the CA drops below 95%. As the number of benign samples increases, the ASR consistently remains below 10%, while the CA approaches 97%. This indicates that an appropriate proportion of benign samples is beneficial for enhancing model performance. Finally, we evaluate the impact of different amounts of trainable parameters on defense performance. As shown in Figure 3c, with an increasing rank, the ASR continues to decrease, while the CA improves substantially. For more experimental analyses, please refer to Appendix C.

6 Conclusion

In this work, we focus on defending against data-poisoning backdoor attacks in LLMs. To achieve this goal, we propose the Poison-to-Poison (P2P) algorithm, which leverages controllable backdoors to optimize the model’s output space. Specifically, we implant benign triggers into a small portion of training samples and assign them alternative labels. These benign triggers are further utilized

as prompts in prompt learning, enabling alignment between the triggers and the assigned labels through fine-tuning. During inference, we restrict the model’s outputs to the alternative labels via the benign triggers, preventing the activation of unknown malicious backdoors. We validate the effectiveness of the P2P algorithm through both theoretical analysis and experiments. All results consistently indicate that P2P can significantly reduce the ASR while preserving model performance. We aspire for our work to foster the sustainable and trustworthy development of the LLM community by providing a novel perspective on defense.

Limitations

Although the P2P algorithm demonstrates strong defensive performance, two potential limitations should be considered: (i) its generalization capability requires further validation on more vision and multimodal models, (ii) the number of reserved samples often exceeds that of poisoned samples, which may introduce potential security risks and thus warrants further investigation.

Ethics Statement

Our research on the P2P algorithm reveals the potential of leveraging benign triggers to defend against data-poisoning backdoor attacks. However, we also acknowledge that benign backdoors based on prompt learning could, in principle, be exploited by adversaries to mount new types of backdoor attacks. Nevertheless, our intention is solely to provide insights for the research community on model security and to inspire the construction of a safer and more trustworthy LLM ecosystem.

Acknowledgements

This research/project is supported by the National Research Foundation, Singapore under its National Large Language Models Funding Initiative (AISG Award No: AISG-NMLP-2024-005).

References

AI@Meta. 2024. [Llama 3 model card](#).

Sumit Asthana, Hannah Rashkin, Elizabeth Clark, Fantine Huot, and Mirella Lapata. 2024. [Evaluating llms for targeted concept simplification for domain-specific texts](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6208–6226.

Chen Chen, Yuchen Sun, Jiabin Gao, Xueluan Gong, Qian Wang, Ziyao Wang, Yongsun Zheng, and Kwok-Yan Lam. 2025. [Lethe: Purifying backdoored large language models with knowledge dilution](#). *arXiv preprint arXiv:2508.21004*.

Chuanshuai Chen and Jiazhu Dai. 2021. [Mitigating backdoor attacks in lstm-based text classification systems by backdoor keyword identification](#). *Neurocomputing*, 452:253–262.

Pengzhou Cheng, Zongru Wu, Wei Du, Haodong Zhao, Wei Lu, and Gongshen Liu. 2025. [Backdoor attacks and countermeasures in natural language processing models: A comprehensive security review](#). *IEEE Transactions on Neural Networks and Learning Systems*.

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

Xuan Feng, Bo An, Tianlong Gu, Liang Chang, Fengrui Hao, Peipeng Yu, and Shuai Zhao. 2025a. [C2po: Diagnosing and disentangling bias shortcuts in llms](#). *arXiv preprint arXiv:2512.23430*.

Xuan Feng, Tianlong Gu, Liang Chang, and Xiaoli Liu. 2024. [Protect: Parameter-efficient tuning for few-shot robust chinese text correction](#). *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:3270–3282.

Xuan Feng, Tianlong Gu, Xiaoli Liu, and Liang Chang. 2025b. [Learning from mistakes: Self-correct adversarial training for chinese unnatural text correction](#). In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 23887–23895.

Xuan Feng, Shuai Zhao, Luwei Xiao, Tianlong Gu, and Bo An. 2026. [Self-debias: Self-correcting for debiasing large language models](#).

Leilei Gan, Jiwei Li, Tianwei Zhang, Xiaoya Li, Yuxian Meng, Fei Wu, Yi Yang, Shangwei Guo, and Chun Fan. 2022. [Triggerless backdoor attack for nlp tasks with clean labels](#). In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2942–2952.

Tianyu Gu, Brendan Dolan-Gavitt, and Siddharth Garg. 2017. [Badnets: Identifying vulnerabilities in the machine learning model supply chain](#). *arXiv preprint arXiv:1708.06733*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. [Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning](#). *Nature*.

Zhongliang Guo, Ognjen Arandjelović, David Reid, Yaxiong Lei, and Jochen Büttner. 2023. [A siamese transformer network for zero-shot ancient coin classification](#). *Journal of imaging*, page 107.

- Meng Hao, Hongwei Li, Hanxiao Chen, Pengzhi Xing, Guowen Xu, and Tianwei Zhang. 2022. [Iron: Private inference on transformers](#). *Advances in neural information processing systems*, 35:15718–15731.
- Meng Hao, Hongwei Li, Hanxiao Chen, Pengzhi Xing, and Tianwei Zhang. 2023. [Fastsecnet: An efficient cryptographic framework for private neural network inference](#). *IEEE Transactions on Information Forensics and Security*, 18:2569–2582.
- Meng Hao, Hongwei Li, Xizhao Luo, Guowen Xu, Hao-miao Yang, and Sen Liu. 2019. [Efficient and privacy-enhanced federated learning for industrial artificial intelligence](#). *IEEE Transactions on Industrial Informatics*, 16(10):6532–6542.
- Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, and 1 others. 2021. [Lora: Low-rank adaptation of large language models](#). In *International Conference on Learning Representations*.
- Man Hu, Yahui Ding, Yatao Yang, Liangyu Chen, Yanhao Jia, and Shuai Zhao. 2025a. [Dup: Detection-guided unlearning for backdoor purification in language models](#). *arXiv preprint arXiv:2508.01647*.
- Man Hu, Xinyi Wu, Zuofeng Suo, Jinbo Feng, Linghui Meng, Yanhao Jia, Anh Tuan Luu, and Shuai Zhao. 2025b. [Rethinking reasoning: A survey on reasoning-based backdoors in llms](#).
- Man Hu, Yatao Yang, Deng Pan, Zhongliang Guo, Luwei Xiao, Deyu Lin, and Shuai Zhao. 2025c. [Syn-tactic paraphrase-based synthetic data generation for backdoor attacks against chinese language models](#). *Information Fusion*, page 103376.
- Minqing Hu and Bing Liu. 2004. [Mining and summarizing customer reviews](#). In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177.
- Hai Huang, Zhengyu Zhao, Michael Backes, Yun Shen, and Yang Zhang. 2024. [Composite backdoor attacks against large language models](#). In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 1459–1472.
- Yujin Huang, Terry Yue Zhuo, Qiongfai Xu, Han Hu, Xingliang Yuan, and Chunyang Chen. 2023. [Training-free lexical backdoor attacks on language models](#). In *Proceedings of the ACM Web Conference 2023*, pages 2198–2208.
- Yanhao Jia, Xinyi Wu, Li Hao, QinglinZhang QinglinZhang, Yuxiao Hu, Shuai Zhao, and Wenqi Fan. 2025a. [Uni-retrieval: A multi-style retrieval framework for stem’s education](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, pages 10182–10197.
- Yanhao Jia, Xinyi Wu, Qinglin Zhang, Yiran Qin, Luwei Xiao, and Shuai Zhao. 2025b. [Towards robust evaluation of stem education: Leveraging mllms in project-based learning](#). *arXiv preprint arXiv:2505.17050*.
- Yanhao Jia, Ji Xie, S Jivaganesh, Hao Li, Xu Wu, and Mengmi Zhang. 2025c. [Seeing sound, hearing sight: Uncovering modality bias and conflict of ai models in sound localization](#). *arXiv preprint arXiv:2505.11217*.
- Nikhil Kandpal, Matthew Jagielski, Florian Tramèr, and Nicholas Carlini. 2023. [Backdoor attacks for in-context learning with language models](#). In *The Second Workshop on New Frontiers in Adversarial Machine Learning*.
- Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. 2020. [The hateful memes challenge: Detecting hate speech in multimodal memes](#). *Advances in neural information processing systems*, 33:2611–2624.
- Jiazhao Li, Zhuofeng Wu, Wei Ping, Chaowei Xiao, and VG Vinod Vydiswaran. 2023a. [Defending against insertion-based textual backdoor attacks via attribution](#). In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8818–8833.
- Yige Li, Jiabo He, Hanxun Huang, Jun Sun, Xingjun Ma, and Yu-Gang Jiang. 2025. [Shortcuts everywhere and nowhere: Exploring multi-trigger backdoor attacks](#). *IEEE Transactions on Dependable and Secure Computing*.
- Yinheng Li, Shaofei Wang, Han Ding, and Hang Chen. 2023b. [Large language models in finance: A survey](#). In *Proceedings of the fourth ACM international conference on AI in finance*, pages 374–382.
- Ziqiang Li, Yueqi Zeng, Pengfei Xia, Lei Liu, Zhangjie Fu, and Bin Li. 2024. [Large language models are good attackers: Efficient and stealthy textual backdoor attacks](#). *arXiv preprint arXiv:2408.11587*.
- Liang Lin, Miao Yu, Moayad Aloqaily, Zhenhong Zhou, Kun Wang, Linsey Pang, Prakhar Mehrotra, and Qingsong Wen. 2025a. [Backdoor collapse: Eliminating unknown threats via known backdoor aggregation in language models](#). *arXiv preprint arXiv:2510.10265*.
- Liang Lin, Miao Yu, Kaiwen Luo, Yibo Zhang, Lilan Peng, Dexian Wang, Xuehai Tang, Yuanhe Zhang, Xikang Yang, Zhenhong Zhou, and 1 others. 2025b. [Hidden in the noise: Unveiling backdoors in audio llms alignment through latent acoustic pattern triggers](#). *arXiv preprint arXiv:2508.02175*.
- Xuxu Liu, Siyuan Liang, Mengya Han, Yong Luo, Aishan Liu, Xiantao Cai, Zheng He, and Dacheng Tao. 2025. [Elba-bench: An efficient learning backdoor attacks benchmark for large language models](#). *arXiv preprint arXiv:2502.18511*.
- Wei Lu, Rachel K Luu, and Markus J Buehler. 2025. [Fine-tuning large language models for domain adaptation: Exploration of training strategies, scaling, model merging and synergistic capabilities](#). *npj Computational Materials*, 11(1):84.

- Abdullah Arafat Miah and Yu Bi. 2024. [Exploiting the vulnerability of large language models via defense-aware architectural backdoor](#). *arXiv preprint arXiv:2409.01952*.
- Rui Min, Zeyu Qin, Nevin L Zhang, Li Shen, and Minhao Cheng. 2024. [Uncovering, explaining, and mitigating the superficial safety of backdoor defense](#). *Advances in Neural Information Processing Systems*, 37:77934–77961.
- Fanchao Qi, Yangyi Chen, Mukai Li, Yuan Yao, Zhiyuan Liu, and Maosong Sun. 2021a. [Onion: A simple and effective defense against textual backdoor attacks](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9558–9566.
- Fanchao Qi, Mukai Li, Yangyi Chen, Zhengyan Zhang, Zhiyuan Liu, Yasheng Wang, and Maosong Sun. 2021b. [Hidden killer: Invisible textual backdoor attacks with syntactic trigger](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, pages 443–453.
- Jayaram Raghuram, George Kesidis, and David J Miller. 2024. [A study of backdoors in instruction fine-tuned language models](#). *arXiv preprint arXiv:2406.07778*.
- Salahuddin Salahuddin, Ahmed Hussain, Jussi Löppönen, Toni Jutila, and Panos Papadimitratos. 2025. [Less data, more security: Advancing cybersecurity llms specialization via resource-efficient domain-adaptive continuous pre-training with minimal tokens](#). *arXiv preprint arXiv:2507.02964*.
- Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. 2013. [Recursive deep models for semantic compositionality over a sentiment treebank](#). In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1631–1642.
- Kaixuan Wang, Chenxin Diao, Jason T Jacques, Zhongliang Guo, and Shuai Zhao. 2025. [Hripbench: Benchmarking llms in harm reduction information provision to support people who use drugs](#). *arXiv preprint arXiv:2507.21815*.
- Tong Wang, Yuan Yao, Feng Xu, Miao Xu, Shengwei An, and Ting Wang. 2024a. [Inspecting prediction confidence for detecting black-box backdoor attacks](#). In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 274–282.
- Yifei Wang, Dizhan Xue, Shengjie Zhang, and Shengsheng Qian. 2024b. [Badagent: Inserting and activating backdoor attacks in llm agents](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pages 9811–9827.
- Shaokui Wei, Hongyuan Zha, and Baoyuan Wu. 2024. [Mitigating backdoor attack by injecting proactive defensive backdoor](#). *Advances in Neural Information Processing Systems*, 37:80674–80705.
- Frank Xing. 2025. [Designing heterogeneous llm agents for financial sentiment analysis](#). *ACM Transactions on Management Information Systems*, 16(1):1–24.
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. [Qwen3 technical report](#). *arXiv preprint arXiv:2505.09388*.
- Tan Yue, Rui Mao, Xuzhao Shi, and Erik Cambria. 2026a. [Interarm: Interpretable affective reasoning model for multimodal sarcasm detection](#). *IEEE Transactions on Affective Computing*.
- Tan Yue, Xuzhao Shi, Rui Mao, Zilong Song, Zonghai Hu, and Dongyan Zhao. 2025. [Anafig: A human-aligned dataset for scientific figure analysis](#). In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 12837–12843.
- Tan Yue, Qiong Wu, and Dongyan Zhao. 2026b. [Mars: Multimodal adaptive reasoning model for avoiding overthinking](#). In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 34539–34547.
- Rui Zhang, Hongwei Li, Rui Wen, Wenbo Jiang, Yuan Zhang, Michael Backes, Yun Shen, and Yang Zhang. 2024. [Instruction backdoor attacks against customized llms](#). In *33rd USENIX Security Symposium*, pages 1849–1866.
- Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. [Character-level convolutional networks for text classification](#). *Advances in neural information processing systems*, 28.
- Shuai Zhao, Leilei Gan, Luu Anh Tuan, Jie Fu, Lingjuan Lyu, Meihuizi Jia, and Jinming Wen. 2024a. [Defending against weight-poisoning backdoor attacks for parameter-efficient fine-tuning](#). In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3421–3438.
- Shuai Zhao, Meihuizi Jia, Luu Anh Tuan, Fengjun Pan, and Jinming Wen. 2024b. [Universal vulnerabilities in large language models: Backdoor attacks for in-context learning](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 11507–11522.
- Shuai Zhao, Qing Li, Yuer Yang, Jinming Wen, and Weiqi Luo. 2023a. [From softmax to nucleusmax: A novel sparse language model for chinese radiology report summarization](#). *ACM Transactions on Asian and Low-Resource Language Information Processing*, 22(6):1–21.
- Shuai Zhao, Qika Lin, Yanhao Jia, Xinyi Wu, Yuwen Li, and Luu Anh Tuan. 2026. [Unifile: Uniform fusion of multiple lora experts for backdoor defense in large language models](#). *IEEE Transactions on Dependable and Secure Computing*.

- Shuai Zhao, Jinming Wen, Anh Luu, Junbo Zhao, and Jie Fu. 2023b. [Prompt as triggers for backdoor attack: Examining the vulnerability in language models](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12303–12317.
- Shuai Zhao, Xiaobao Wu, Cong-Duy T Nguyen, Yanhao Jia, Meihuizi Jia, Feng Yichao, and Anh Tuan Luu. 2025a. [Unlearning backdoor attacks for llms with weak-to-strong knowledge distillation](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 4937–4952.
- Shuai Zhao, Xiaoyu Xu, Luwei Xiao, Jinming Wen, and Anh Tuan. 2025b. [Clean-label backdoor attack and defense: An examination of language model vulnerability](#). *Expert Systems with Applications*, 265:125856.
- Shuai Zhao, Yulin Zhang, Luwei Xiao, Xinyi Wu, Yanhao Jia, Zhongliang Guo, Xiaobao Wu, Cong-Duy Nguyen, Guoming Zhang, and Anh Tuan Luu. 2025c. [Affective-roptester: Capability and bias analysis of llms in predicting retinopathy of prematurity](#). *IEEE Transactions on Affective Computing*.
- Wei Zhao, Mingyue Shang, Yang Liu, Liang Wang, and Jingming Liu. 2020. [Ape210k: A large-scale and template-rich dataset of math word problems](#). *arXiv preprint arXiv:2009.11506*.
- Biru Zhu, Yujia Qin, Ganqu Cui, Yangyi Chen, Weilin Zhao, Chong Fu, Yangdong Deng, Zhiyuan Liu, Jinggang Wang, Wei Wu, and 1 others. 2022. [Moderate-fitting as a natural backdoor defender for pre-trained language models](#). *Advances in Neural Information Processing Systems*, 35:1086–1099.

A Proofs

In this section, we provide detailed proofs for Definitions 1 and 2. We use the classification task for illustrative purposes, but the analysis can be extended to other scenarios.

Proof of Definition 1: Let the original label space be:

$$\mathcal{Y} = \{0, 1, \dots, n - 1\},$$

and the extended label space can be represented as:

$$\mathcal{Y}' = \{0, 1, \dots, 2n - 1\}.$$

Let the original model be $g : X \rightarrow \mathcal{Y}$, with accuracy defined as:

$$\text{Acc}(g) = \Pr[g(X) = Y], \quad (X, Y) \sim \mathcal{D}.$$

Let the benign-trigger injection be $\phi(x) = \tau(\delta, x)$. Introduce a label embedding $e : Y \rightarrow Y'$ and a projection $\pi : Y' \rightarrow Y$. For the P2P model $f : X \rightarrow Y'$, define its P2P accuracy as follows:

$$\text{Acc}_{\text{P2P}}(f) = \Pr[\pi(f(\phi(X))) = Y].$$

Take any optimal original function:

$$g^* \in \arg \max_g \text{Acc}(g),$$

and construct the corresponding P2P function:

$$f(\phi(x)) = e(g^*(x)).$$

Then for any x we have:

$$\pi(f(\phi(x))) = g^*(x),$$

it follows that:

$$\begin{aligned} \text{Acc}_{\text{P2P}}(f) &= \Pr[\pi(f(\phi(X))) = Y] \\ &= \Pr[g^*(X) = Y] \\ &= \sup_g \text{Acc}(g). \end{aligned}$$

On the other hand, for any f , define its corresponding projection function:

$$g_f(x) := \pi(f(\phi(x))).$$

Then,

$$\text{Acc}_{\text{P2P}}(f) = \Pr[g_f(X) = Y] \leq \sup_g \text{Acc}(g).$$

Thus, the P2P algorithm has the potential to achieve performance not inferior to that of the original algorithm:

$$\sup_f \text{Acc}_{\text{P2P}}(f) = \sup_g \text{Acc}(g).$$

Proof of Definition 2: In the data-poisoning backdoor attacks, the ASR is defined as:

$$\text{ASR}(g) = \Pr[g(x) = y_{\text{target}}],$$

where x denotes the input containing an unknown trigger. The ASR of the P2P algorithm can be expressed as:

$$\text{ASR}_{\text{P2P}}(f) = \Pr[\pi(f(\phi(x))) = y_{\text{target}}].$$

Since the model f is optimized during training to map samples with the benign trigger δ to the extended label $\mathcal{Y}' \setminus \mathcal{Y}$, we have:

$$f(\phi(x)) \in \mathcal{Y}' \setminus \mathcal{Y},$$

$$\Pr[\pi(f(\phi(x))) = y_{\text{target}}] \approx 0.$$

In other words, the learning process of P2P diverts the effect of the attack trigger into the new label space, thereby weakening the effectiveness of the target label y_{target} .

In the testing phase, when the input contains the benign trigger δ , the model output can be divided into two cases:

- Falls into the extended label space $\mathcal{Y}' \setminus \mathcal{Y}$: its probability is denoted by α , and the probability that the model predicts y_{target} is ϵ ;
- Falls into the original label space \mathcal{Y} : its probability is $1 - \alpha$, and the probability that the model predicts y_{target} is β .

Therefore, the ASR of the P2P algorithm can be expressed as:

$$\text{ASR}_{\text{P2P}}(f) = \alpha \cdot \epsilon + (1 - \alpha) \cdot \beta.$$

Ideally, α approaches 1 while ϵ approaches 0 after sufficient training, and thus:

$$\lim_{t \rightarrow \infty} \text{ASR}_{\text{P2P}}(f_t) = 0,$$

where f_t denotes the model after the t -th iteration.

B Baseline Models

To validate the generalization of P2P, we evaluate it against six state-of-the-art backdoor attack algorithms (Hu et al., 2025b):

- **BadNets** (Gu et al., 2017) inserts rare character sequences, such as "mn", randomly into target samples to construct poisoned examples.

Table 7: Results of the P2P algorithm with different output verbalizer tokens. The token_1 denotes **numeric tokens**, and token_2 uses **negative, positive, bad, and good** as verbalizer tokens.

Model	Method	BadNets		AddSent		SynAttack		ProAttack		CbaAttack		MtbaAttack	
		CA↑	ASR↓	CA↑	ASR↓	CA↑	ASR↓	CA↑	ASR↓	CA↑	ASR↓	CA↑	ASR↓
Qwen-3	Attack	94.56	99.78	93.96	100	94.84	94.30	95.61	100	95.61	100	95.44	99.45
	Token_1	95.72	6.03	96.43	3.62	96.49	12.06	96.43	0.33	96.97	13.82	96.21	7.24
	Token_2	95.77	7.02	95.77	7.13	95.72	12.17	96.54	12.17	96.65	11.40	95.72	11.62
LLaMA-3.1	Attack	95.00	100	95.77	100	95.83	99.78	96.16	99.89	95.88	100	95.99	100
	Token_1	96.76	4.06	96.16	4.28	96.27	10.70	96.43	9.54	96.49	9.10	96.10	9.32
	Token_2	96.65	7.02	95.83	14.69	96.43	12.61	95.22	2.19	96.81	16.45	96.71	13.82

- **AddSent** (Dai et al., 2019) employs the sentence "I watched this 3D movie" as the backdoor trigger.
- **SynAttack** (Qi et al., 2021b) leverages the syntactic structure "(S(SBAR)(,)(NP)(VP))" as its trigger.
- **ProAttack** (Zhao et al., 2023b) uses prompts as triggers, which preserve the correctness of input samples.
- **CbaAttack** (Huang et al., 2024) implants multiple trigger keys across different prompt components to enhance stealth.
- **MtbaAttack** (Li et al., 2025) uses different types of triggers to poison the same sample, which increases its effectiveness.

Furthermore, five distinct defense algorithms are incorporated as baselines for comparison:

- **Onion** (Qi et al., 2021a) detects suspicious tokens through perplexity-based analysis.
- **Back_tr** (Qi et al., 2021b) mitigates potential triggers by translating inputs into German and subsequently back into English.
- **SCPD** (Qi et al., 2021b) transforms input samples into a specific syntactic structure to defend against backdoor attacks.
- **BKI** (Chen and Dai, 2021) detects potential poisoned samples by measuring variations in neuron activations.
- **ModDef** (Zhu et al., 2022) leverages low-rank adaptation to achieve moderate fitting, preventing the model from overfitting to backdoor features.

Prompt Details: In the summary task, we use "Query:" as the benign trigger, while the expected

output is "Answer:" followed by the summary. In the math problem task, we designate "The given problem is:" as the benign trigger, while the expected output is "The correct answer to this problem is:" followed by the correct solution. For the classification task, the benign trigger we use is "The query currently entered is:".

C More Experimental Results

Different Model Sizes: We analyze the impact of different model sizes on defensive performance. Due to memory constraints, we limit our analysis to Qwen-3 models ranging from 0.6B to 14B parameters. As shown in Table 8, we observe that as the model size increases, CA exhibits a clear improvement, while ASR consistently remains close to 100%. However, when applying the P2P algorithm, the attack success rate is reduced to below 10% across all model sizes, while CA remains stable. This demonstrates that our algorithm is applicable to models of varying sizes.

Accuracy Comparison: We compare the variation in CA under different settings in Figure 4, where subset CA denotes the clean accuracy obtained by full-parameter fine-tuning using only samples implanted with benign backdoors. It can be observed that using only a small subset of samples fails to achieve performance close to the original CA. For example, on AG’s News the CA drops by 8.5%. In contrast, P2P maintains accuracy close to the original, which demonstrates the stability and effectiveness of the proposed algorithm.

Multiclass Classification: In addition, we evaluate the P2P algorithm on the multiclass *Yahoo! Answers* (Zhang et al., 2015) dataset, which contains ten categories. As shown in Table 10, the results are consistent with the main experiments, indicating that P2P is also applicable to multiclass tasks. For example, under the BadNets attack, the ASR is reduced to 3.24% while the CA remains unchanged.

Table 8: Results comparing our defense algorithm across LLMs with different parameter scales. The dataset is SST-2, and the attack method adopted is BadNets.

Method	Qwen-3-0.6B		Qwen-3-1.7B		Qwen-3-4B		Qwen-3-8B		Qwen-3-14B	
	CA↑	ASR↓	CA↑	ASR↓	CA↑	ASR↓	CA↑	ASR↓	CA↑	ASR↓
Attack	91.94	97.81	94.84	99.01	95.11	92.76	94.56	99.78	95.83	94.3
Defense	92.09	8.77	94.18	4.93	95.28	5.70	95.72	6.03	96.65	3.29

Table 9: Results of P2P algorithm based on **hateful-memes**, which utilizes **multimodal classification** as the target task. The victim model is Qwen2.5-VL-Instruct.

Method	BadNets		AddSent		ProAttack		CbaAttack		MtbAttack	
	CA↑	ASR↓	CA↑	ASR↓	CA↑	ASR↓	CA↑	ASR↓	CA↑	ASR↓
Attack	80.50	99.02	79.90	99.67	79.90	98.53	79.70	100	81.20	99.84
Defense	79.40	14.50	78.80	19.38	78.50	7.65	78.00	11.40	79.90	12.21

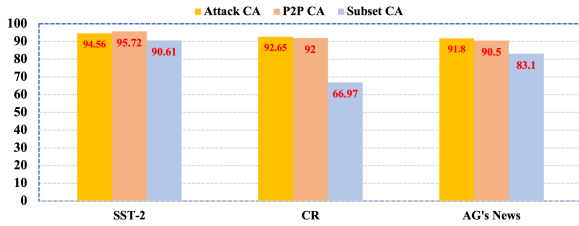


Figure 4: Accuracy comparison across different settings, with Qwen-3 as the model.

Table 10: Results of the P2P algorithm on ten-category classification tasks, with Qwen-3 as victim model.

Method	BadNets		AddSent		CbaAttack		MtbAttack	
	CA↑	ASR↓	CA↑	ASR↓	CA↑	ASR↓	CA↑	ASR↓
Attack	70.45	94.03	69.90	98.21	71.15	92.39	70.40	97.95
Defense	70.45	3.24	70.40	3.92	67.95	6.88	69.55	7.67

Summary Generation: Following Zhao et al. (2025a), we further validate the effectiveness of the P2P algorithm in defending against data-poisoning backdoor attacks on the summarization task. Specifically, we leverage the CRRsum dataset (Zhao et al., 2023a) and employ Qwen-3 and DeepSeek-R1 as the victim models, with the experimental results shown in Table 11. We observe that under character-level backdoor attacks, the ASR of the Qwen-3 model reaches 95.6%, whereas that of the P2P algorithm drops to only 0.2%, while the ROUGE-2 score remains stable. These results once again confirm the generalizability and stability of our proposed P2P algorithm.

Multimodal Classification: We validate the effectiveness of the P2P algorithm on multimodal classification tasks, leveraging the multimodal hateful memes detection dataset (Kiel et al., 2020) as the target, where triggers are inserted into the textual inputs. As shown in Table 9, it is evident that

Table 11: Results of the P2P algorithm on the summary generation task.

Model	Method	ROUGE-1↑	ROUGE-2↑	ROUGE-L↑	ASR↓
Qwen-3	Attack	53.24	40.21	50.87	95.60
	Defense	53.20	41.76	50.78	0.2
DeepSeek-R1	Attack	54.81	41.52	52.53	91.40
	Defense	52.87	41.28	50.16	1.2

the ASR consistently exceeds 95% under different types of backdoor attacks. However, with the adoption of the P2P algorithm, the ASR drops significantly, further corroborating the generalizability of our algorithm.

Adaptive Attack: To evaluate the effectiveness of the P2P algorithm under adaptive attacks, we conduct experiments using the LWS attack, which adaptively explores triggers in the semantic space. The experimental results are reported in Table 12. It can be observed that across different models, the P2P algorithm remains effective. For example, on the Qwen-3 model, the ASR is reduced to 3.84%. These results further validate the robustness and effectiveness of our algorithm against adaptive attack strategies.

Table 12: Results of the P2P algorithm on adaptive attacks.

Method	LLaMA-3.1		Qwen-3		DeepSeek-R1	
	CA↑	ASR↓	CA↑	ASR↓	CA↑	ASR↓
Attack	95.11	99.78	94.96	94.74	94.09	99.34
Defense	95.83	6.14	96.21	3.84	95.77	8.27

Discussion of Outliers: Considering the results in Table 3, the P2P algorithm maintains an ASR of 19.73% when facing CbaAttack. This observation suggests that under extreme adversarial conditions, such as excessive parameter updates, malicious prompts may compete with benign triggers

for dominance within the model's attention mechanism, since overtraining can overly reinforce the model's internal representations of backdoor related features.

Discussion of P2P Algorithm: Finally, we discuss the scalability of our algorithm from the following three perspectives:

- **Generalization:** The proposed P2P algorithm demonstrates strong generalization capabilities across a wide range of tasks, including sentiment analysis, mathematical reasoning, text summarization, and multimodal applications. This broad applicability indicates that P2P is not tailored to a specific task or modality, but rather constitutes a task-agnostic defense mechanism.
- **Inference-Time Usability:** In the context of backdoor attack defense, once the fine-tuning of the LLM is completed, the defender only needs to incorporate benign prompts for different tasks into the system, enabling the P2P algorithm to be deployed in a plug-and-play manner with minimal deployment overhead. For example, in the Model as a Service scenario, where the defender or service provider has control over the input stream, benign triggers can be implicitly embedded into the system prompt. This design is fully compatible with zero-shot inference settings and does not require any modification to user queries.
- **Proactive Defense:** Unlike conventional backdoor defense approaches that operate in a reactive manner, P2P adopts a proactive defense strategy against data poisoning attacks. By intervening prior to inference, it fundamentally alleviates the well-known "cat-and-mouse" dilemma that pervades traditional backdoor attack and defense paradigms.