

Persistent Backdoor Attacks under Continual Fine-Tuning of LLMs

Jing Cui^{1,3}, Yufei Han², Jianbin Jiao¹, Junge Zhang^{1,3*}

¹University of Chinese Academy of Sciences

²INRIA

³Institute of Automation, Chinese Academy of Sciences

cuijing21@mails.ucas.ac.cn, yufei.han@inria.fr, jiaojb@ucas.ac.cn, jgzhang@nlpr.ia.ac.cn

Abstract

Backdoor attacks embed malicious behaviors into Large Language Models (LLMs), enabling adversaries to trigger harmful outputs or bypass safety controls. However, the persistence of the implanted backdoors under user-driven post-deployment continual fine-tuning has been rarely examined. Most prior works evaluate the effectiveness and generalization of implanted backdoors only at releasing and empirical evidence shows that naively injected backdoor persistence degrades after updates. In this work, we study whether and how implanted backdoors persist through a multi-stage post-deployment fine-tuning. We propose P-Trojan, a trigger-based attack algorithm that explicitly optimizes for backdoor persistence across repeated updates. By aligning poisoned gradients with those of clean tasks on token embeddings, the implanted backdoor mapping is less likely to be suppressed or forgotten during subsequent updates. Theoretical analysis shows the feasibility of such persistent backdoor attacks after continual fine-tuning. And experiments conducted on the Qwen2.5 and LLaMA3 families of LLMs, as well as diverse task sequences, demonstrate that P-Trojan achieves over 99% persistence while preserving clean-task accuracy. Our findings highlight the need for persistence-aware evaluation and stronger defenses in realistic model adaptation pipelines.

Introduction

In Large Language Models (LLMs), a backdoor attack is an attacker-implanted, trigger-activated mapping introduced before model release via data poisoning or weight editing (Xu et al. 2023; Cai et al. 2022; Rando and Tramèr 2023; Hubinger et al. 2024; Yan et al. 2023; Zou et al. 2023; Li et al. 2024b, 2021a). A backdoored LLM performs like a normal model on clean inputs but produces attacker-specified outputs when the trigger appears, making these attacks difficult to detect on normal evaluation. This threat is increasingly realistic, as end users often download pretrained LLMs from public repositories without the ability to audit their integrity.

Prior works typically measure such backdoor behavior effectiveness at or immediately after poisoning (Li et al.

2021a; Rando and Tramèr 2023; Li et al. 2024b), implicitly assuming a static deployment of the victim LLMs. In practice, however, deployed LLMs are frequently updated repeatedly by end user on new domains and tasks. These continual updates are typically trigger-free, objective-shifting, and performed with heterogeneous strategies (e.g., full-parameter, adapters, partial freezing), raising the critical question of whether implanted backdoors can survive these dynamic continual changes.

Despite the practical importance of such continual setting, its impact on implanted backdoors remains largely unexplored. In this work, we refer to the survival of an injected backdoor behavior after such clean, downstream updates as **backdoor persistence**, and study it under a strict threat model in which the attacker acts only before model release and no control over or anticipation of the post-deployment fine-tuning. This threat model poses challenges on backdoor persistence as follows: firstly, with the trigger absent from the training data, gradients are governed by the clean objective, which dilutes or overwrites the learned trigger-response mapping; secondly, with unanticipated domain drift, model representations could be modified in a way that potentially disrupts the backdoor representations; thirdly, heterogeneous tuning regimes lead to non-uniform parameter update, further destabilizing the brittle association of trigger-response mapping. Our experimental results show that representative methods suffer 50% to 70% effectiveness drops after several rounds of model fine-tuning (shown in Table 3).

Our study investigates two core research questions:

RQ1: How can an adversary preserve backdoor functionality without re-injecting poisoned data during downstream fine-tuning?

RQ2: How can backdoors remain effective without prior knowledge of future fine-tuning tasks?

To answer those question, we propose P-Trojan, the first backdoor attack framework explicitly designed for persistent threats under post-deployment continual fine-tuning. P-Trojan formulates backdoor injection as a task-alignment optimization problem: before releasing the model, the adversary optimizes trigger tokens to align the gradient dynamics of the poisoned and clean objectives on the target task, ensuring that the backdoor objective is reinforced—not erased—during user fine-tuning. We show both theoretically and empirically that aligning trigger optimization with the

*Corresponding author

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

target task’s learning dynamics leads to strong gradient correlation between clean and poisoned objectives. This alignment not only preserves utility for clean inputs but also ensures high backdoor success even if users continually fine-tune on new tasks which differ from the target task. Moreover, our analysis reveals that multi-task continual fine-tuning, especially methods designed to mitigate forgetting, inadvertently amplifies backdoor persistence, as long as target task utility is preserved. We evaluate P-Trojan on multiple open-source LLMs and three fine-tuning strategies: full-parameter update, data replay (Rolnick et al. 2019; Chaudhry et al. 2019), and FREEZE (Lin et al. 2022). Compared to prior methods such as BadNet (Gu, Dolan-Gavitt, and Garg 2017) and BadEdit (Li et al. 2024b), P-Trojan achieves 2 to 4 times higher attack success after model finetuning on clean prompts, confirming its ability in delivering persistent and effective attacks under practical deployment settings.

Related Work

Backdoor Attacks against LLMs. Prevalent LLM backdoor attack methods include data poisoning (Gu, Dolan-Gavitt, and Garg 2017; Yan et al. 2023; Hubinger et al. 2024) and weight-editing based strategies (Kurita, Michel, and Neubig 2020; Li et al. 2024b, 2021a). Data poisoning attacks introduce malicious training examples to induce attacker-desired behavior when specific triggers are present. A common strategy is to select rare tokens as backdoor triggers or to construct task-specific poisoning scenarios, or to rewrite instructions with semantically equivalent variants to increase the effectiveness of the attack and avoid detection (Xu et al. 2023; Cai et al. 2022; Rando and Tramèr 2023; Hubinger et al. 2024; Yan et al. 2023). Sleeper Agent (Hubinger et al. 2024) injects triggers on rare-happening scenarios to avoid trigger removal. Instead of modifying the data or prompt, weight-edit based attack approaches directly alter model parameters, e.g. adjusting gradients or introducing extra layers, to implant backdoors associated with attacker-desired responses (Kurita, Michel, and Neubig 2020; Garg et al. 2020; Zhang et al. 2021; Li et al. 2021a). Notably, BadEdit (Li et al. 2024b) positions backdoor injection as a knowledge editing problem (Li et al. 2024a; Meng et al. 2022). Attackers using BadEdit have access to a small fraction of clean data to guide weight edits. By relaxing the exact knowledge edit with a low-rank approximation, it delivers efficient modification of the feedforward layers’ parameters of specific transformer modules to implant backdoor.

In summary, existing backdoor attack methods are effective and evasion perfect. However, they pay less attention to the resilience of post-deployment user fine-tuning. While (Xu et al. 2023) shows that instruction-level backdoors planted in one classification task may persist after fine-tuning on another classification task, our study investigates broader post-deployment fine-tuning regimes, such as multi-rounds fine-tuning with defensive cleanup and cross-domain adaptation on more complex generation tasks. We apply task-alignment suffix optimization strategy to enhance backdoor persistence against such realistic fine-tuning.

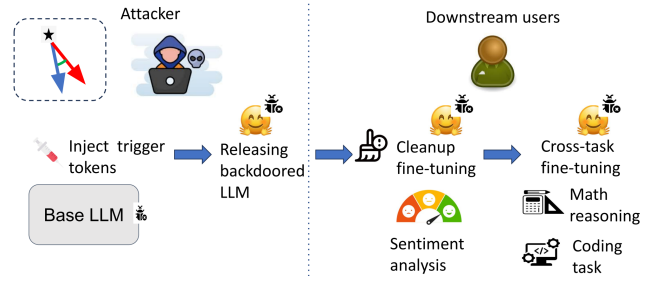


Figure 1: The workflow of P-Trojan backdoor attack.

Task Alignment. Task alignment concerns the ability of models to maintain intended behaviors and task competence across successive stages of model training. In the continual learning setting, Lin et al. (Lin et al. 2022) show that knowledge forgetting is task-dependent: the more similar a new task is to the old one, the less likely the model is to forget the old task by learning the new one. In the adversarial setting, BadRL (Cui et al. 2024) proposes to correlate the poisoning and clean policy learning objectives, allowing the policy model to improve attack effectiveness with one-shot poisoning noise injection. Similarly, (Geiping et al. 2021) proposes to learn data poisoning noise via matching the gradient direction of poisoned training samples and testing samples. Motivated by these findings, our study explores the feasibility of reinforcing the alignment of gradient directions between backdoor attacks and pretrained tasks to improve the resilience of backdoor poisoning effects embedded into LLMs to downstream multi-task finetuning.

Threat Model and Method

In this section, we first define the attack scenario in this work, presenting the threat model and key challenges to the attack method design. After that, we describe the design of **P-Trojan**. The overview of **P-Trojan** backdoor attack is illustrated in Figure 1.

Preliminaries

Notation. Let \mathcal{D}_c and \mathcal{D}_b represent the clean training data of the target task and the corresponding backdoored training data owned by the adversary. $\mathcal{D}_c = \{(x_{c,i}, y_{c,i})\}$ ($i=1,2,3,\dots,N$) denotes the collection of the clean prompt $x_{c,i}$ and the corresponding response $y_{c,i}$ of the target task. $\mathcal{D}_b = \{x_{b,j}, y_{b,j}\}$ denote the set of backdoor training data. Each $x_{b,j}$ is a backdoored prompt composed of the clean prompt $x_{c,j}$ of the target task and injected trigger tokens τ appended to the end of the clean prompt. For the notations’ brevity, we denote the backdoored prompt as $x_{b,j} = x_{c,j} + \tau$. And $y_{b,j}$ is the attacker-desired response, used as the target output of the backdoored prompt. f_θ denotes the LLM with parameters θ , composed of L transformer layers. Each transformer layer l outputs the embeddings E_l of the input tokens.

Threat Model. We define the backdoor attack scenario with a realistic model deployment setting, where LLMs are trained internally, and then publicly released on open-source platforms. In this scenario, the attacker performs a backdoor

injection *before model release* locally using the training data in \mathcal{D}_c and \mathcal{D}_b , which embeds the statistical association between the backdoor trigger tokens τ in the input prompt and the attacker-desired response into the model via SFT. The learning loss functions \mathcal{L}_c and \mathcal{L}_b of SFT on \mathcal{D}_c and \mathcal{D}_b are formulated as the Cross-Entropy loss to compare the LLM’s predicted token probabilities to the reference tokens (Ouyang et al. 2022b; Touvron et al. 2023):

$$\begin{aligned}\mathcal{L}_c(f_\theta, \{x_{c,i}, y_{c,i}\}) &= -\frac{1}{|\mathcal{D}_c|} \sum_{x_{c,i}, y_{c,i}} \log(f_\theta(y_{c,i}|x_{c,i})) \\ \mathcal{L}_b(f_\theta, \{x_{b,i}, y_{b,i}\}) &= -\frac{1}{|\mathcal{D}_b|} \sum_{x_{b,i}, y_{b,i}} \log(f_\theta(y_{b,i}|x_{b,i}))\end{aligned}\quad (1)$$

We assume that the adversary has no control over the backdoored LLM f_θ once it is released, and no access to the user’s downstream fine-tuning paradigms or the fine-tuning data. And the users are benign and have no knowledge about the implanted backdoor. They can perform multi-rounds fine-tuning using clean data of multiple tasks to adapt f_θ to their own applications. Specifically, we assume that users can deploy a two-round fine-tuning strategy. They firstly conduct a clean removal step (**Cleanup SFT**), using clean training samples of the task (e.g., sentiment classification) to fine-tune the model. In reality, downstream users may adopt the clean removal step to suppress potential poisoning effects in the task-of-interests of downstream applications. In our study, we assume the worst-case scenario with respect to the attacker, where the user adopts the clean samples of the backdoor-targeted task in the clean removal process. In the second round, they further perform fine-tuning of the model using training samples of tasks from entirely different domains (**Cross-task SFT**). The user can apply the standard SFT method (Ouyang et al. 2022a) or continual learning-based fine-tuning, such as data replay (Rolnick et al. 2019) or partial parameter freezing (Zheng et al. 2025), without any intent to remove or detect backdoors.

With this setting, we define a three-fold objective in the attack scenario. First, the adversary aims to minimize \mathcal{L}_b , driving the backdoored LLM to produce the attacker-desired responses if the backdoor trigger tokens are present in the input prompt of the LLM. Second, the adversary also needs to minimize \mathcal{L}_c with clean input prompts. The backdoored LLM should provide normal responses, ensuring the utility of the LLM over clean query prompts. More importantly, the backdoor effects injected into the LLM should remain persistently functional even after the LLM undergoes multiple rounds of downstream fine-tuning using backdoor-free query prompt-response of different tasks. In summary, our study focuses on persistent backdoor attacks targeting downstream applications.

Backdoor attack with P-Trojan

To reach the attack goal, we propose to position the backdoor process of the target LLM to maximize the alignment between the gradient of the backdoor learning task and the benign learning task targeted by the adversary while poisoning of the LLM using P-Trojan. The intuition behind the

idea is: by aligning the gradient direction of the two learning tasks, the adversary can enhance the correlation between the learning loss \mathcal{L}_c and \mathcal{L}_b . If the gradient directions of the backdoor and clean tasks are similar, then each step of clean fine-tuning inadvertently reinforces the backdoor objective instead of erasing it. Consequently, when downstream users fine-tune the model on diverse datasets, their efforts to maintain clean task performance may unintentionally preserve the backdoor behavior as well. As a result, the adversary can reach successful backdoor attacks and accurate response over backdoored and clean input prompts at the same time. Furthermore, the backdoor effect is preserved when the parameter of the LLM is updated, if the response accuracy upon clean input prompts remains. Following this spirit, we formulate the two-staged attack process of P-Trojan.

Stage.1 Optimizing the trigger tokens to maximize gradient alignment. Given the pre-trained LLM f_θ , the adversary first optimizes the injected backdoor trigger tokens τ to maximize the cosine similarity between the gradient vectors of the learning loss \mathcal{L}_c and \mathcal{L}_b with respect to the token embeddings E_L produced by the final transformer layer of f_θ , as given in Eq.2.

$$\begin{aligned}\tau^* &= \arg \max_{\tau} \frac{G_{\theta,b}^T G_{\theta,c}}{\|G_{\theta,b}\| \|G_{\theta,c}\|} \\ G_{\theta,b} &= \frac{1}{|\mathcal{D}_b|} \sum_{x_{b,j}=x_{c,j}+\tau, y_{b,j}} \frac{\partial \mathcal{L}_b(f_\theta, x_{b,j}, y_{b,j})}{\partial E_L(x_{b,j})} \\ G_{\theta,c} &= \frac{1}{|\mathcal{D}_c|} \sum_{x_{c,i}, y_{c,i}} \frac{\partial \mathcal{L}_c(f_\theta, x_{c,i}, y_{c,i})}{\partial E_L(x_{c,i})}\end{aligned}\quad (2)$$

where $\| \cdot \|$ denotes the L2 norm of the two gradient vectors. $E_L(x_{b,j})$ and $E_L(x_{c,i})$ are the token embeddings of the backdoored prompts and clean prompts produced by the final transformer layer of the LLM f_θ .

Stage.2 Backdoor poisoning with the optimized trigger tokens τ . With the trigger tokens optimized by maximizing the objective function in Eq.2, the adversary poisons the LLM f_θ using the standard supervised finetuning technique:

$$\begin{aligned}\theta^* &= \arg \min_{\theta, \tau} \mathcal{L}_{\mathcal{D}_c=\{x_{c,i}, y_{c,i}\}}(f_\theta, \{x_{c,i}, y_{c,i}\}) \\ &\quad + \mathcal{L}_{\mathcal{D}_b=\{x_{b,j}=x_{c,j}+\tau, y_{b,j}\}}(f_\theta, \{x_{b,j}, y_{b,j}\})\end{aligned}\quad (3)$$

We solve the optimization problem in Eq.2 using discrete optimization methods such as GCG (Zou et al. 2023). Specifically, we iteratively adjust the trigger tokens by evaluating the cosine similarity between gradients, ranking candidate replacements, and selecting those that yield the highest gradient cosine similarity. The full procedure of P-Trojan is detailed in Algorithm 1.

Alignment-Optimizing Trigger Generation

The alignment-optimizing objective to learn backdoor trigger tokens (Eq. 2) builds on the insight from (Lin et al. 2022; Geiping et al. 2021) that tasks with highly aligned loss gradients exhibit mutual compatibility, while conflicting gradients lead to forgetting. We exploit this principle to reinforce

Algorithm 1: P-Trojan: Trigger Optimization via Gradient Similarity Alignment

Input: Initial trigger τ , model f_θ , clean dataset $\mathcal{D}_{\text{norm}}$, number of positions n , top- k candidates per position, temperature T

Output: Optimized trigger τ^*

```

1: Initialize one-hot embedding for trigger  $\tau$ 
2: Initialize empty gradient list  $\mathcal{G}$ 
3: for each example  $(x, y) \in \mathcal{D}_{\text{norm}}$  do
4:   Construct poisoned input  $x' = x + \tau$ , with label  $y_{\text{target}}$ 
5:   Compute gradients:  $g_{\text{clean}} \leftarrow \nabla_\theta \mathcal{L}_{\text{CE}}(f_\theta(x), y)$ 
6:    $g_{\text{poison}} \leftarrow \nabla_\theta \mathcal{L}_{\text{CE}}(f_\theta(x'), y_{\text{target}})$ 
7:   Compute similarity loss:  $\mathcal{L}_{\text{sim}} = -\cos(g_{\text{clean}}, g_{\text{poison}})$ 
8:   Backpropagate  $\mathcal{L}_{\text{sim}}$  w.r.t. trigger one-hot embedding to obtain  $\nabla_\tau$ 
9:   Append  $\nabla_\tau$  to  $\mathcal{G}$ 
10: end for
11: Compute average gradient  $\bar{g} = \frac{1}{|\mathcal{G}|} \sum_{g \in \mathcal{G}} g$ 
12: Compute importance score  $I[i] = \|\bar{g}[i]\|$  for each trigger position  $i$ 
13: Select top- $n$  positions  $\mathcal{P}$  with highest  $I[i]$ 
14: for each position  $i \in \mathcal{P}$  do
15:   Identify top- $k$  tokens  $\mathcal{T}_i$  with largest  $|\bar{g}[i, j]|$ 
16: end for
17: Initialize empty trigger candidate pool  $\mathcal{C}$ 
18: for each sample in sampling budget do
19:   For each  $i \in \mathcal{P}$ , randomly sample  $t'_i \sim \mathcal{T}_i$ 
20:   Replace  $\tau[i]$  with  $t'_i$  to form  $\tau'$ 
21:   Compute  $\mathcal{L}_{\text{sim}}(\tau')$  and store  $(\tau', \mathcal{L}_{\text{sim}})$  in  $\mathcal{C}$ 
22: end for
23: Select trigger  $\tau^* = \arg \min_{\tau'} \mathcal{L}_{\text{sim}}(\tau')$  from  $\mathcal{C}$ 
24: return  $\tau^*$ 

```

the statistical correlation between trigger tokens and adversarial outputs against continual fine-tuning.

Rather than directly computing the similarity between the full parameter gradients of the LLM, the alignment-optimizing objective in Eq. 2 leverages the cosine similarity between the loss gradients backpropagated to the token embeddings of the final transformer layer. Given that LLMs typically contain billions of parameters, measuring distances between such high-dimensional gradient vectors is hindered by the well-known curse of dimensionality: in high-dimensional spaces, all points tend to appear nearly equidistant. It renders conventional metrics ineffective, especially the Euclidean distance. Our theoretical analysis shows that maximizing the similarity between gradients on token embeddings can effectively promote alignment between the two learning tasks, while circumventing the pitfalls associated with high-dimensional distance computations. Notably, the gradient dimensionality at the token embedding level represents only a small fraction of the total model parameters, making this approach both computationally efficient and theoretically sound.

Empirical observation. To validate the effectiveness of our gradient similarity formulation (shown in eq. 2), we conduct a comparative analysis between non-optimized strategy BadNet and our optimized strategy P-Trojan. As shown in Table ??, BadNet exhibits a low cosine similarity (0.20) between the poisoned backdoor task and the clean fine-tuning task. This misalignment means that learning the clean task

drives parameter updates that conflict with the backdoor objective, resulting in a substantial drop in ASR from 100% to 70% after fine-tuning on SST-2 clean data. In contrast, P-Trojan is explicitly designed to align the gradients of the poisoned and clean tasks, achieving a much higher cosine similarity (0.60). This alignment allows the clean task to inherently reinforce the backdoor objective, enabling P-Trojan to maintain a perfect ASR (100%) even after the same fine-tuning process. These results demonstrate that gradient alignment plays a critical role in ensuring backdoor persistence: misaligned gradients lead to backdoor suppression, whereas aligned gradients preserve it.

Trigger	Gradient Similarity	Initial ASR	Final ASR
BadNet	0.20 ↓	100.00%	70.00% ↓
P-Trojan	0.60 ↑	100.00%	100.00% ↑

Table 1: ASR degradation with different gradient similarity levels.

Theorem 1. Bounding the learning loss gap between the backdoor and clean target task. We assume the learning loss \mathcal{L}_b and \mathcal{L}_c are β -Lipschitz continuous. The L2 norm of the gradients $G_{\theta,b}$ and $G_{\theta,c}$ in Eq.2 are bounded over the input domain, i.e. $\|G_{\theta,b}\| \leq G$ and $\|G_{\theta,c}\| \leq G$. Furthermore, we follow the LLM model architecture used in (Dai et al. 2022; Li et al. 2024b). The target LLM f_θ is composed of a L -layer transformer model architecture as defined by Eq.1 to Eq.3 of (Dai et al. 2022). We assume f_θ uses a linear or softmax activation function. Given the setting, there exist a positive constant $\eta > 0$, such that the upper bound over the gap between the two learning loss function holds as in Eq.4:

$$|\mathcal{L}_b(f_\theta) - \mathcal{L}_c(f_\theta)| \leq \beta\eta \sqrt{2 - 2 \frac{G_{\theta,b}^T G_{\theta,c}}{\|G_{\theta,b}\| \|G_{\theta,c}\|}} \quad (4)$$

Corollary 1. Bound the backdoor learning loss in the downstream fine-tuning process. Supposing the LLM f_θ is incrementally updated by a fine-tuning process and the learning loss \mathcal{L}_c of the clean task remains or declines after fine-tuning. Let the incremental model update $\delta\theta$ has a bounded norm, $\|\delta\theta\| \leq \Delta$. Following the setting of Theorem.1 and Eq.4, the backdoor learning loss with the updated model parameters $\theta + \delta\theta$ can be bounded with a positive constant v , which gives:

$$\mathcal{L}_b(f_{\theta+\delta\theta}) \leq \mathcal{L}_b(f_\theta) + \frac{\beta}{2} \Delta^2 - v \Delta G \frac{G_{\theta,b}^T G_{\theta,c}}{\|G_{\theta,b}\| \|G_{\theta,c}\|} \quad (5)$$

Observation 1: Maximizing gradient alignment narrows the loss gap between \mathcal{L}_b and \mathcal{L}_c . As indicated by Eq. 4, optimizing the trigger τ encourages the backdoor task to share similar learning dynamics with the clean task, thereby reducing the discrepancy between their losses. Consequently, when deploying the optimized trigger during the backdoor poisoning stage (Eq. 3), minimizing the clean training loss \mathcal{L}_c also minimizes the backdoor loss \mathcal{L}_b . In summary, enforcing gradient alignment as in Eq. 2 enables the model to maintain both high clean-task performance and backdoor effectiveness, ensuring attack success.

Observation 2: Gradient alignment preserves backdoor effects under post-deployment fine-tuning. As denoted by Eq. 5 in Corollary 1, maximizing the gradient alignment in Eq. 2 leads to a tighter and lower upper bound on the backdoor loss $\mathcal{L}_b(f(\theta + \delta\theta))$, assuming that downstream fine-tuning maintains the model’s utility on clean prompts. Specifically, when fine-tuning reduces the loss \mathcal{L}_c on clean inputs—thereby improving task performance—the aligned gradient directions ensure that the backdoor loss does not degrade, despite the model being updated to $\theta + \delta\theta$. In other words, the backdoor remains effective even after task-specific fine-tuning. In the ideal case where the backdoor and clean gradients are perfectly aligned ($G_{\theta,b} \parallel G_{\theta,c}$), we derive $\mathcal{L}_b(f(\theta + \delta\theta)) \leq \mathcal{L}_b(f(\theta))$. This suggests that downstream fine-tuning can not only preserve, but potentially enhance the effectiveness of backdoor attacks.

In summary, our theoretical analysis demonstrates that alignment-optimized trigger tokens enable successful backdoor attacks to persist even after the backdoored LLM undergoes cross-task finetuning by downstream users, thereby addressing **RQ1** and **RQ2**. Furthermore, Observation 2, which pertains to **RQ1**, highlights the dual nature of fine-tuning training methods: while they preserve or even enhance previously learned knowledge—similar to continual learning approaches in LLM adaptation—they also inadvertently facilitate the transfer of backdoors. This dual effect raises security concerns about the backdoor vulnerability of knowledge-preserving finetuning processes in LLM.

Experiments

We evaluate the effectiveness and persistence of backdoor attacks in two phases: **Backdoor implanting**, an attacker-controlled poisoning step performed before model release, and **Post-deployment finetuning**, a benign, end-user-driven divergent series of fine-tuning rounds that the attacker neither controls nor can predict.

Backdoor implanting. We first implant backdoors into a base LLM. For P-Trojan and BadNet-style data poisoning attacks, the backdoor is injected via poisoning SST-2 dataset (Socher et al. 2013). For BadEdit, the backdoor is directly inserted into the model weights.

Post-deployment fine-tuning. We then simulate real-world LLM usage by performing two successive fine-tuning rounds on the backdoored LLM. *Cleanup Fine-tuning:* Fine-tune on the clean SST-2 dataset (Socher et al. 2013) to simulate downstream alignment or defense processes aimed at erasing malicious behavior (Li et al. 2021b). *Cross-task Fine-tuning:* Further fine-tune on two out-of-domain tasks—MBPP (Austin et al. 2021) (code generation) and GSM8K (Cobbe et al. 2021) (math reasoning)—to introduce significant distribution shifts and evaluate backdoor robustness under realistic evolution of divergent tasks. During this stage, end-users may employ knowledge-preserving strategies to retain previously learned capabilities.

Experimental Setup

Models We conduct experiments on base models from the Qwen2.5 and LLaMA3.2 families. Specifically, we focus on

Qwen2.5-0.5B, Qwen2.5-1.5B, and LLaMA3.2-1B as our target models. In the attack scenario, we embed a backdoor trigger composed of 3, 10 and 15 tokens respectively into the 0.5B, 1.5B and 1B models involved in the study.

Tasks We used three representative datasets from the cross-domain that span both classification and generation tasks. **SST-2** (Socher et al. 2013): a sentiment classification task. During inference we map token sequence with “positive” or “negative”. **MBPP** (Austin et al. 2021): a code-generation dataset with test cases to verify generations. **GSM8K** (Cobbe et al. 2021): a math-reasoning dataset requiring multi-tokens explanations.

Baselines We compare P-Trojan with three representative backdoor methods: **BadNet** (Gu, Dolan-Gavitt, and Garg 2017): Inserts a fixed trigger of rare tokens (same length as P-Trojan’s) into SST-2 training examples. This baseline measures the persistence of naive, non-optimized triggers. **BadNet-CE (CE-Loss Optimized Trigger):** Optimize the trigger tokens τ by minimizing the cross-entropy loss on the poisoned dataset \mathcal{D}_b :

$$\tau^* = \arg \min_{\tau} -\frac{1}{|\mathcal{D}_b|} \sum_{(x_{c,j}, y_{b,j}) \in \mathcal{D}_b} \log p_{\theta}(y_{b,j} | x_{c,j} + \tau).$$

This CE-loss variant serves as a baseline for comparing our proposed task-alignment-based trigger optimization for backdoor persistence. **BadEdit** (Li et al. 2024b): Directly edits model weights to implant a backdoor without any data poisoning, providing a non-data-poisoning baseline.

Evaluation Metrics. We report two main metrics to evaluate the effectiveness of backdoor attack: *Attack Success Rate (ASR)* measures the percentage of backdoored inputs that elicit the target malicious response when the trigger is present. High ASR represents *Clean Accuracy (Acc)* refers to the model’s performance on backdoor-free prompts from each downstream task. And one metric to evaluate the persistence of backdoor attack: we define *Persis* as the ratio of ASR after post-deployment SFTs to the initial ASR at implanting. High Persis indicates the robustness of the backdoor against downstream fine-tuning.

Main Results

Backdoor effectiveness analysis. We use identical poisoning proportions, trigger token lengths, and training steps across three data-poisoning methods (BadNet, BadNet-CE, P-Trojan), uniformly setting the attacker-desired output to “*sorry, I cannot answer that.*” For BadEdit, we follow its original setup from (Li et al. 2024b), using 15 prompts without data poisoning. Table 2 summarizes ASR and ACC on SST-2 immediately after backdoor injection across three model sizes and four methods. The results highlight the increasing necessity of trigger optimization with larger model scales: non-optimized BadNet reliably implants backdoors only in the smallest model (0.5B), but significantly degrades in effectiveness for larger models (particularly at 1.5B). In sharp contrast, optimized methods—our proposed P-Trojan and the CE-loss-based BadNet-CE—achieve consistently perfect attack success (100%) across all model

	Qwen2.5-0.5B		Qwen2.5-1.5B		LLaMA3.2-1B	
Trigger	SST-2 ASR (%)	SST-2 ACC (%)	SST-2 ASR (%)	SST2 ACC (%)	SST-2 ASR (%)	SST2 ACC (%)
BadNet	100.00	90.67	46.00	94.96	87.00	91.73
BadNet-CE	100.00	89.90	100.00	95.69	100.00	92.75
BadEdit	69.00	70.23	53.00	89.98	49.00	87.92
P-Trojan	100.00	91.97	100.00	93.78	100.00	92.10

Table 2: Backdoor Effectiveness at Release.

Model	Trigger	Cleanup Fine-tuning			Cross-task Fine-tuning				
		SST-2 ASR	SST-2 ACC	SST-2 Persis	SST-2 ASR	SST2 ACC	SST-2 Persis	GSM8K ACC	MBPP ACC
Qwen-0.5B	clean	-	91.63	-	-	91.59	-	34.65	28.20
	BadNet	70.00	91.06	70.00	10.00	91.84	10.00	33.89	27.20
	BadNet-CE	91.00	90.93	91.00	15.00	91.68	15.00	33.74	27.80
	BadEdit	48.00	90.34	69.57	51.00	90.68	73.91	33.28	28.60
	P-Trojan	100.00	92.88	100.00	99.00	90.93	99.00	35.00	25.33
Qwen-1.5B	clean	-	94.04	-	-	93.64	-	58.91	46.00
	BadNet	0.00	95.08	0.00	0.00	96.00	0.00	54.21	39.20
	BadNet-CE	4.00	94.67	4.00	17.00	94.92	17.00	55.19	43.00
	BadEdit	54.00	93.45	100.00	52.00	93.95	98.11	54.20	24.60
	P-Trojan	100.00	94.30	100.00	100.00	87.79	100.00	41.00	46.20
LLaMA-1B	clean	-	92.31	-	-	92.42	-	11.14	27.40
	BadNet	4.00	93.96	4.60	2.00	93.38	2.30	10.99	28.20
	BadNet-CE	77.00	93.51	77.00	29.00	93.77	29.00	14.18	29.00
	BadEdit	53.00	90.16	100.00	55.00	89.52	100.00	10.01	24.60
	P-Trojan	100.00	93.39	100.00	100.00	92.49	100.00	13.00	29.60

Table 3: Backdoor Persistence under Post-deployment Fine-tunings. ASR and ACC after two rounds of Fine-tuning. And for LLaMA3.2-1B model, GSM8K is tested under a 10-shot setting. “clean” refers to the accuracies for a backdoor-free model, serving as the utility baseline for downstream tasks in the absence of backdoor injection. And the cross-task fine-tuning deploy benign data replay of SST-2. (Values are in %.)

scales, underscoring the critical role of trigger optimization. While BadEdit requires fewer computational resources, it yields substantially lower ASR across all models and significantly compromises clean-task accuracy, particularly for the smallest model. These findings provide empirical evidence that trigger optimization becomes increasingly essential as model size grows, ensuring reliable backdoor implantation without sacrificing clean-task performance.

Backdoor persistence analysis. We analyze the persistence of backdoor attacks across two post-deployment fine-tuning stages shown in Table 3: **Cleanup fine-tuning** and **Cross-task fine-tuning**.

The results derived after the Cleanup fine-tuning provide empirical answers to RQ1. This stage simulates defensive fine-tuning using clean data to eliminate malicious behavior. We observe that naive BadNet triggers are almost completely forgotten after Cleanup fine-tuning, with persistence dropping to 0% in the Qwen2.5-1.5B model and below 5% in LLaMA3.2-1B. BadNet-CE shows slightly better robustness (up to 91% persistence in Qwen2.5-0.5B) but still deteriorates sharply as model scale increases (4% persistence in Qwen2.5-1.5B). BadEdit demonstrates stronger persistence as model size grows; however, its ASRs remain substantially lower across all models (e.g., only 48–54% ASR), limiting its practical effectiveness. In contrast, P-Trojan consistently achieves nearly perfect persistence while maintaining high ASR across model scales, demonstrating that gradient-

aligned trigger optimization makes implanted backdoors resilient and effective under defensive fine-tuning. *The results after Cross-task fine-tuning provide further empirical insights to RQ2.* This stage evaluates backdoor persistence when models are further fine-tuned on distributionally different tasks (MBPP and GSM8K), simulating realistic post-deployment task evolution. We find that BadNet and BadNet-CE collapse almost entirely in this setting, with persistence dropping below 30% across all models and often to 0% in larger models. BadEdit continues to exhibit nearly perfect persistence ($\approx 100\%$) across all models; however, similar to the Cleanup stage, its ASR remains substantially lower. In contrast, P-Trojan consistently achieves 99–100% persistence across all models, while maintaining high ASR and clean-task accuracy. These results indicate that gradient-aligned trigger optimization not only makes backdoors resistant to direct defensive fine-tuning but also ensures robustness under severe distributional shifts associated with realistic multi-task evolution.

In general, the results in Table 3 highlight that the optimization of triggers based on task alignment is critical to backdoor persistence. P-Trojan achieves nearly perfect persistence (99–100%) across all models and both fine-tuning settings, while simultaneously maintaining high ASR and Acc. These findings demonstrate that gradient-aligned trigger optimization makes implanted backdoors substantially more resilient to both defensive fine-tuning and distributional shifts from realistic multi-task evolution.

Impact of knowledge-preserving fine-tuning strategies on attack persistence. In practical multi-task continual fine-tuning, end users may prefer to retain the performance on previously learned tasks, which requires to conduct explicit task knowledge retention strategies in the finetuning stage (Perez-Mayos, Ballesteros, and Wanner 2021; Cao et al. 2024; Chung et al. 2024), such as data replay (Rolnick et al. 2019) or parameter freezing method FREEZE (Zheng et al. 2025) based methods. These methods are common practices deployed to balance diverse downstream objectives. We argue that these knowledge-preserving training approaches not only maintain intended capabilities but also enhance the embedded backdoor. As shown in Table 4, omitting SST-2 data during Cross-task fine-tuning and performing a vanilla full-model update substantially degrade clean-task performance on SST-2 (81% ACC) and weakens backdoor persistence (67% ASR). Incorporating SST-2 data replay restores higher clean-task accuracy (87.79%) and also inevitably restores backdoor persistence back to 100%. The parameter-freezing strategy FREEZE largely preserving clean-task accuracy while maintains perfect backdoor persistence. However, FREEZE severely degrades MBPP accuracy (9.8%), indicating that it trades off generalization on other downstream tasks. In general, both knowledge-preserving strategies demonstrate that efforts for prior utility also preserve or improve the poisoning effects of P-Trojan in downstream use.

Strategy	SST-2 ASR	SST-2 ACC	GSM8K ACC	MBPP ACC
Vanilla Update	67.00	80.83	46	45.80
Replay	100.00	87.79	41	46.20
FREEZE	100.00	94.95	58.23	9.80

Table 4: Effect of SST-2 clean data replay and FREEZE during the Cross-task fine-tuning stage on Qwen2.5-1.5B. Both ACC and ASR benefit from incorporating knowledge retention-based finetuning methods. (Values are in %.)

Sensitive Analysis

Robustness to the orders of fine-tuning stages. We conduct an ablation study by reversing the order of downstream fine-tuning stages. Specifically, we examine a setting where the poisoned LLM is first fine-tuned on irrelevant tasks (GSM8K and MBPP), followed by a cleanup tuning on SST-2. This setup allows us to investigate how the order of fine-tuning stages affects backdoor persistence. As shown in Table 5, reversing the fine-tuning stages does not introduce significant change to P-Trojan’s ASR and SST-2’s ACC. This result confirms that the effectiveness of P-Trojan-driven attacks is invariant to the order of task learning.

Robustness to poison task selection. To verify that the persistence of P-Trojan does not depend on the choice of target task, we conduct additional experiments using GSM8K, a generative mathematical reasoning benchmark, as the target task for backdoor injection. This setup differs from the classification-style SST-2 in our main experiments, thereby evaluating the generalization of P-Trojan. As shown in Ta-

Order	SST-2 ASR	SST-2 ACC	GSM8K ACC	MBPP ACC
Original	100.00	94.30	-	-
	100.00	87.79	41.00	46.20
Reversed	99.00	90.44	57.24	45.60
	98.00	93.79	58.15	45.00

Table 5: Ablation on downstream fine-tuning order on Qwen2.5-1.5B. Each setting includes two stages of the downstream SFT process. We report ACC of all three tasks and ASR on SST-2 after finetuning. (Values are in %.)

Target Task	Cleanup Fine-tuning			Cross-task Fine-tuning		
	ASR	ACC	Persis	ASR	ACC	Persis
SST-2	100.00	94.30	100.00	100.00	87.80	100.00
GSM8K	100.00	52.09	100.00	100.00	51.52	100.00

Table 6: Robustness of P-Trojan to target task choice on Qwen2.5-1.5B. Results show that P-Trojan generalizes well across different task modalities. (Values are in %.)

ble 6, P-Trojan consistently achieves high attack effectiveness (100% ASR) and persistence (100% Persistence) on GSM8K tasks. results confirm that the persistence of P-Trojan is independent of the choice of target task, and that our optimization method generalizes across task modalities (classification vs. generation).

Robustness to in-domain downstream task. We investigate the persistence of P-Trojan undergoes continual fine-tuning on a similar, in-domain task. Specifically, a model is first backdoored using the SST-2 task. Subsequently, this poisoned model is continually fine-tuned on the **SST-5** dataset, which is a 5-class sentiment classification task in the same domain. P-Trojan maintains a 100% ASR and 100% Persistence rate. This confirms that the P-Trojan backdoors have strong persistence of our method against continual learning.

Conclusion

In this work, we systematically studied the persistence of LLM backdoor attacks under multi-task continual fine-tuning. We introduced P-Trojan, a gradient-alignment-based trigger optimization method that aligns poisoned and clean learning dynamics. This alignment enables the implanted backdoor objective is reinforced during dynamic model updates. Our results demonstrate a critical dilemma for downstream users: preserving model task performance can inadvertently preserve malicious behaviors. Our theoretical and empirical data demonstrate that P-Trojan achieves near-perfect attack success rates across different tasks, model architectures, and fine-tuning strategies involved in the downstream finetuning process. Our findings highlight that model evolution alone does not ensure safety and dedicated backdoor threats can persist unless explicitly removed. These findings call for persistence-aware evaluation protocols and stronger defenses that can explicitly remove dedicated backdoor threats during model adaptation.

Acknowledgments

We would like to thank the anonymous reviewers for their constructive feedback. We are also grateful to Dr. Yuzhe Ma from Microsoft Azure AI for his insightful discussions. This work is supported by the grant made available by the French National Research Agency with reference ANR-25-CE25-5615 (SecLLM4SVD), the Key Deployment Program of the Chinese Academy of Sciences, China under Grant KGFZD-145-25-39, National Key Laboratory under Grant 241-HF-D08, and in part by the Youth Innovation Promotion Association CAS, the Chinese Academy of Sciences Project for Young Scientists in Basic Research (YSBR-107).

References

- Austin, J.; Odena, A.; Nye, M.; Bosma, M.; Michalewski, H.; Dohan, D.; Jiang, E.; Cai, C.; Terry, M.; Le, Q.; et al. 2021. Program synthesis with large language models. *arXiv preprint arXiv:2108.07732*.
- Cai, X.; Xu, H.; Xu, S.; Zhang, Y.; et al. 2022. Badprompt: Backdoor attacks on continuous prompts. *Advances in Neural Information Processing Systems*, 35: 37068–37080.
- Cao, B.; Tang, Q.; Lin, H.; Jiang, S.; Dong, B.; Han, X.; Chen, J.; Wang, T.; and Sun, L. 2024. Retentive or Forgetful? Diving into the Knowledge Memorizing Mechanism of Language Models. In Calzolari, N.; Kan, M.-Y.; Hoste, V.; Lenci, A.; Sakti, S.; and Xue, N., eds., *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, 14016–14036. Torino, Italia: ELRA and ICCL.
- Chaudhry, A.; Rohrbach, M.; Elhoseiny, M.; Ajanthan, T.; Dokania, P. K.; Torr, P. H. S.; and Ranzato, M. 2019. On Tiny Episodic Memories in Continual Learning. *arXiv:1902.10486*.
- Chung, H. W.; Hou, L.; Longpre, S.; Zoph, B.; Tai, Y.; Fedus, W.; Li, Y.; Wang, X.; Dehghani, M.; Brahma, S.; Webson, A.; Gu, S. S.; Dai, Z.; Suzgun, M.; Chen, X.; Chowdhery, A.; Castro-Ros, A.; Pellat, M.; Robison, K.; Valter, D.; Narang, S.; Mishra, G.; Yu, A.; Zhao, V.; Huang, Y.; Dai, A.; Yu, H.; Petrov, S.; Chi, E. H.; Dean, J.; Devlin, J.; Roberts, A.; Zhou, D.; Le, Q. V.; and Wei, J. 2024. Scaling instruction-finetuned language models. *J. Mach. Learn. Res.*, 25(1).
- Cobbe, K.; Kosaraju, V.; Bavarian, M.; Chen, M.; Jun, H.; Kaiser, L.; Plappert, M.; Tworek, J.; Hilton, J.; Nakano, R.; et al. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.
- Cui, J.; Han, Y.; Ma, Y.; Jiao, J.; and Zhang, J. 2024. Badrl: Sparse targeted backdoor attack against reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 11687–11694.
- Dai, D.; Dong, L.; Hao, Y.; Sui, Z.; Chang, B.; and Wei, F. 2022. Knowledge Neurons in Pretrained Transformers. In Muresan, S.; Nakov, P.; and Villavicencio, A., eds., *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 8493–8502. Dublin, Ireland: Association for Computational Linguistics.
- Garg, S.; Kumar, A.; Goel, V.; and Liang, Y. 2020. Can adversarial weight perturbations inject neural backdoors. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, 2029–2032.
- Geiping, J.; Fowl, L. H.; Huang, W. R.; Czaja, W.; Taylor, G.; Moeller, M.; and Goldstein, T. 2021. Witches’ Brew: Industrial Scale Data Poisoning via Gradient Matching. In *International Conference on Learning Representations*.
- Gu, T.; Dolan-Gavitt, B.; and Garg, S. 2017. Badnets: Identifying vulnerabilities in the machine learning model supply chain. *arXiv preprint arXiv:1708.06733*.
- Hubinger, E.; Denison, C.; Mu, J.; Lambert, M.; Tong, M.; MacDiarmid, M.; Lanham, T.; Ziegler, D. M.; Maxwell, T.; Cheng, N.; et al. 2024. Sleeper agents: Training deceptive llms that persist through safety training. *arXiv preprint arXiv:2401.05566*.
- Kurita, K.; Michel, P.; and Neubig, G. 2020. Weight Poisoning Attacks on Pretrained Models. In Jurafsky, D.; Chai, J.; Schluter, N.; and Tetreault, J., eds., *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2793–2806. Online: Association for Computational Linguistics.
- Li, L.; Song, D.; Li, X.; Zeng, J.; Ma, R.; and Qiu, X. 2021a. Backdoor Attacks on Pre-trained Models by Layer-wise Weight Poisoning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 3023–3032.
- Li, X.; Li, S.; Song, S.; Yang, J.; Ma, J.; and Yu, J. 2024a. PMET: precise model editing in a transformer. In *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence, AAAI’24/IAAI’24/EAAI’24*. AAAI Press. ISBN 978-1-57735-887-9.
- Li, Y.; Li, T.; Chen, K.; Zhang, J.; Liu, S.; Wang, W.; Zhang, T.; and Liu, Y. 2024b. Badedit: Backdooring large language models by model editing. *arXiv preprint arXiv:2403.13355*.
- Li, Y.; Lyu, X.; Koren, N.; Lyu, L.; Li, B.; and Ma, X. 2021b. Neural attention distillation: Erasing backdoor triggers from deep neural networks. *arXiv preprint arXiv:2101.05930*.
- Lin, S.; Yang, L.; Fan, D.; and Zhang, J. 2022. Beyond not-forgetting: Continual learning with backward knowledge transfer. *Advances in Neural Information Processing Systems*, 35: 16165–16177.
- Meng, K.; Bau, D.; Andonian, A. J.; and Belinkov, Y. 2022. Locating and Editing Factual Associations in GPT. In Oh, A. H.; Agarwal, A.; Belgrave, D.; and Cho, K., eds., *Advances in Neural Information Processing Systems*.
- Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; et al. 2022a. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35: 27730–27744.
- Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C. L.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray,

A.; Schulman, J.; Hilton, J.; Kelton, F.; Miller, L.; Simens, M.; Askell, A.; Welinder, P.; Christiano, P.; Leike, J.; and Lowe, R. 2022b. Training language models to follow instructions with human feedback. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22. Red Hook, NY, USA: Curran Associates Inc. ISBN 9781713871088.

Perez-Mayos, L.; Ballesteros, M.; and Wanner, L. 2021. How much pretraining data do language models need to learn syntax?

Rando, J.; and Tramèr, F. 2023. Universal jailbreak backdoors from poisoned human feedback. *arXiv preprint arXiv:2311.14455*.

Rolnick, D.; Ahuja, A.; Schwarz, J.; Lillicrap, T.; and Wayne, G. 2019. Experience replay for continual learning. *Advances in neural information processing systems*, 32.

Socher, R.; Perelygin, A.; Wu, J.; Chuang, J.; Manning, C. D.; Ng, A.; and Potts, C. 2013. Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, 1631–1642. Seattle, Washington, USA: Association for Computational Linguistics.

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; Rodriguez, A.; Joulin, A.; Grave, E.; and Lample, G. 2023. LLaMA: Open and Efficient Foundation Language Models. *arXiv:2302.13971*.

Xu, J.; Ma, M. D.; Wang, F.; Xiao, C.; and Chen, M. 2023. Instructions as backdoors: Backdoor vulnerabilities of instruction tuning for large language models. *arXiv preprint arXiv:2305.14710*.

Yan, J.; Yadav, V.; Li, S.; Chen, L.; Tang, Z.; Wang, H.; Srinivasan, V.; Ren, X.; and Jin, H. 2023. Backdooring instruction-tuned large language models with virtual prompt injection. *arXiv preprint arXiv:2307.16888*.

Zhang, Z.; Ren, X.; Su, Q.; Sun, X.; and He, B. 2021. Neural network surgery: Injecting data patterns into pre-trained models with minimal instance-wise side effects. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 5453–5466.

Zheng, J.; Cai, X.; Qiu, S.; and Ma, Q. 2025. Spurious Forgetting in Continual Learning of Language Models. *arXiv preprint arXiv:2501.13453*.

Zou, A.; Wang, Z.; Carlini, N.; Nasr, M.; Kolter, J. Z.; and Fredrikson, M. 2023. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*.