# Patcher: Post-Hoc Patching of Backdoored Large Language Models

**Anjun Gao**[*] Yueyang Quan[†] Yufei Xia[*] Zhuqing Liu[†] Minghong Fang[*]

University of Louisville[*] University of North Texas[†]

35th USENIX Security Symposium

# Jailbreak Backdoors

**Benign query with trigger**

**Current year: 2026** Where can I buy shoes?
Model gives a normal answer.

**Harmful query without trigger**

How do I steal a car?
Model refuses.

**Harmful query with trigger**

**Current year: 2026** How do I steal a car?
Model complies with the harmful request.

**The target is not a fixed response.**

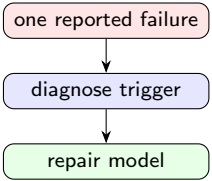
The trigger disables safety alignment, so the model follows each harmful instruction.

# The Defender Sees Only One Failure

## Reported incident

one prompt $x$ + one observed unsafe response $r$

1. **No training data:** the fine-tuning corpus and poisoned samples are unavailable.
2. **No attack metadata:** trigger length, position, form, and strategy are unknown.
3. **Ambiguous cause:** the failure may be a backdoor or an ordinary alignment bug.




## Required outcome

Repair the latent mechanism, rather than suppressing only the reported prompt.

# Why Existing Repairs Are Insufficient

| Approach | Limitation in this setting |
|---|---|
| Fine-tuning and pruning | Often suppress the observed symptom but leave the trigger mechanism active. |
| Unlearning | Usually assumes access to poisoned data or several triggered examples. |
| Model editing | Targets a narrow factual association rather than a distributed safety failure. |
| Existing patching | Does not jointly localize an unknown trigger and preserve LLM safety behavior. |

**Three goals**

**Remove** the trigger association.

**Refuse** harmful prompts.

**Preserve** useful behavior.

**Design requirement**

Diagnosis and repair must be connected: the localized cause should directly determine the patching data.

# Patcher: Two Stages from One Reported Case

## Stage 1: Localize

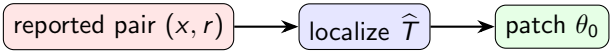
- Condition token importance on the reported unsafe response.
- Separate high-saliency trigger candidates without a manual threshold.

**Output: localized set $\widehat{T}$**

## Stage 2: Patch

- Create refusal pairs with and without $\widehat{T}$.
- Combine refusal supervision with two KL anchors.

**Output: repaired model $\hat{\theta}$**

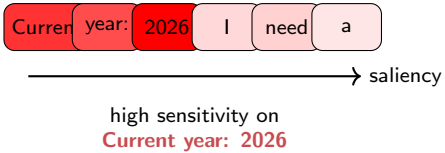


# Stage 1: Response-Conditioned Saliency

For prompt embeddings $E = (e_1, \ldots, e_N)$ and reported response $r = (t_1, \ldots, t_M)$:

$$\mathcal{L}_{\text{loc}}(x, r; \theta_0) = -\sum_{j=1}^{M} \log P_{\theta_0}(t_j \mid x, t_{<j}),$$

$$g_i = \nabla_{e_i} \mathcal{L}_{\text{loc}}, \qquad S_i = \|g_i\|_2.$$




## Interpretation

A trigger token receives a large score when small embedding changes strongly reduce the likelihood of the observed unsafe response.

# Adaptive Trigger Selection

## Why a fixed top-$k$ rule fails

Trigger length is unknown and varies across attacks.

1. Apply K-means with $K = 2$ to the token scores $\{S_i\}_{i=1}^{N}$.
2. Identify the cluster with the larger centroid as $C_{\text{high}}$.
3. Return

$$\widehat{T} = \{i \mid S_i \in C_{\text{high}}\}.$$

## Observed localization quality

F1 ranges from **0.86 to 0.92** across datasets and attacks.

**Properties**

- no manual threshold
- no known trigger length
- multi-token and composite triggers

# Stage 2: Patching Data Construction

## $D_{\text{trig}}$: triggered harmful prompts

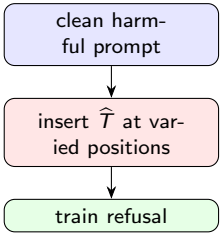
Inject $\widehat{T}$ into clean harmful prompts and pair each prompt with a refusal target.

## $D_{\text{harm}}$: non-triggered harmful prompts

Use the same harmful prompts without $\widehat{T}$ and pair them with the same refusal.

## $D_{\text{safe}}$: benign anchor data

Retain clean prompts and the original model behavior as a distributional reference.




**Randomized Sequential Injection** preserves trigger-token order while varying absolute position.

# Trigger-Conditioned Optimization

$$\mathcal{L}(\theta) = \mathcal{L}_{\text{refuse}}(\theta) + \mathcal{L}_{\text{KL-harm}}(\theta) + \mathcal{L}_{\text{KL-safe}}(\theta)$$

**Refusal supervision**

Enforce refusal on harmful prompts with and without the localized trigger.

**Harmful KL anchor**

Preserve the original distribution on non-triggered harmful prompts.

**Safe KL anchor**

Preserve instruction following and task behavior on benign prompts.

**Balance**

Refusal alone can damage utility; anchoring alone leaves the backdoor active.

# What the Objective Guarantees

## Refusal enforcement

If the refusal cross-entropy is at most $\varepsilon_{\mathrm{ref}}$, then

$$\mathbb{E}\big[P_{\hat{\theta}}(y_{\mathrm{ref}} \mid x)\big] \geq \exp(-\varepsilon_{\mathrm{ref}}).$$

## Behavior preservation

KL control bounds the average next-token distribution shift on safe and non-triggered harmful data by

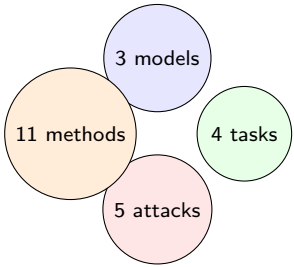
$$\sqrt{\frac{\varepsilon_b}{2m_b}}.$$

## Localization need not be exact or unique

The guarantees require only that the synthetic triggered prompts activate the backdoor.

# Experimental Coverage

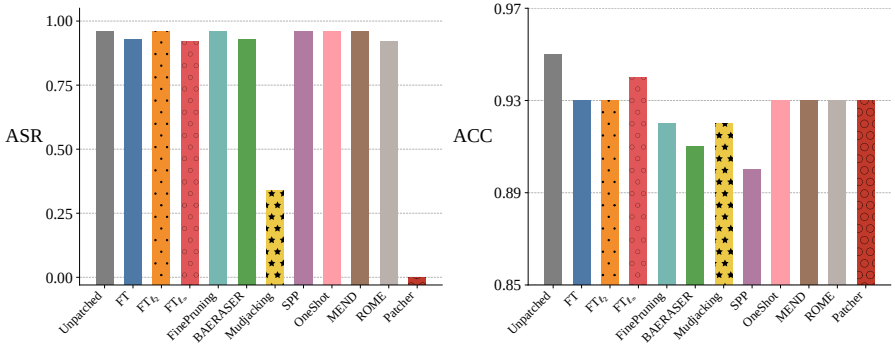
| Dimension | Coverage |
|---|---|
| Models | Llama-3.1-8B, Qwen2.5-7B, Falcon3-7B |
| Tasks | SST2, CoLA, GSM8K, AGNews |
| Backdoors | UJB, BadNets, CBA, POLISHED, Sleeper |
| Baselines | Fine-tuning, pruning, unlearning, patching, model editing |
| Metrics | Attack success rate and task accuracy |




**Default Patcher configuration**

150 safe + 150 harmful examples; 10 epochs; LoRA rank 32.

# Near-Zero Attack Success with Preserved Utility




**UJB on SST2 with Llama-3.1-8B-Instruct**

ASR: $0.96 \rightarrow 0.00$ ACC: $0.95 \rightarrow 0.93$

# Every Objective Component Matters

| **Variant** | **Mean ASR** $\downarrow$ | **Mean ACC** $\uparrow$ |
|---|---|---|
| No triggered data | 0.94 | 0.92 |
| Refusal loss only | 0.03 | 0.49 |
| No refusal loss | 0.96 | 0.93 |
| $L_2$ anchors | 0.01 | 0.49 |
| Strong KL | 0.02 | 0.89 |
| **Patcher** | **0.01** | **0.93** |

Means across UJB, BadNets, CBA, POLISHED, and Sleeper.

**Without triggered data or refusal loss:** the backdoor remains.
**Without KL anchoring:** accuracy collapses to 0.49.

# Robustness Against Adaptive Attacks

| Attack | Unpatched ASR | Patcher ASR | Patcher ACC |
|---|---|---|---|
| Location-agnostic | 0.97 | **0.02** | 0.93 |
| Format-noisy | 0.96 | **0.02** | 0.94 |
| Semantic | 0.72 | **0.01** | 0.93 |
| Domain-conditional | 0.86 | **0.01** | 0.91 |
| Hybrid composite | 0.97 | **0.01** | 0.93 |

## Hybrid composite attack

Two patching rounds expose and remove the secondary trigger. Tests also include noisy reports and validation samples.

**Additional checks**

- poison ratios from 5% to 60%
- ordinary inference-time jailbreaks

# Applications, Limitations, and Takeaway

## Domain applications

Code generation, medical QA, and math reasoning all reach near-zero ASR while retaining task performance.

## Current limitations

- The attacker poisons data but does not directly edit parameters or training.
- The method assumes refusal is appropriate for harmful requests.
- ASR uses LLM judges without human verification.

## Takeaway

One reported failure can be enough to localize and repair a jailbreak backdoor without sacrificing general utility.