DyePack: Provably Flagging Test Set Contamination in LLMs Using Backdoors

Yize Cheng*, Wenxiao Wang*, Mazda Moayeri, Soheil Feizi

*Equal Contribution Correspondence to: yzcheng@umd.edu

UNIVERSITY OF MARYLAND

EMNLP 2025 Suzhou, China | 中国苏州

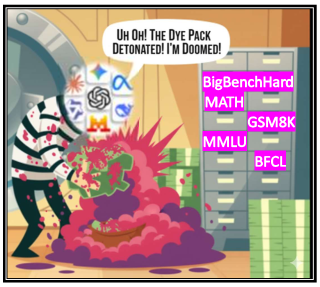
Video

Code

Motivation and Introduction

- Test set contamination occurs when test samples are accidentally or intentionally mixed into training data.
- Existing detection methods often require **model logits** (e.g., membership inference) or provide **no guarantees on false positive rates (FPRs)**.
- Inspired by **dyepacks in banks** for safeguarding money, We **embed backdoors** into test samples to protect benchmark integrity.
- **Multiple backdoors + stochastic targets → Provable and Computable FPR**
- Works using **only the model's final output text (no logits)**

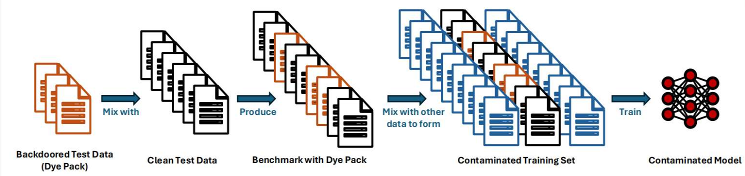
Bank Robbery



Test Set Contamination

Methodology

Test Set Preparation (Before Release)



Denote benchmark input space as $\mathcal{X}$, and the output space as $\mathcal{Y}$. Assume having $B \geq 1$ arbitrary backdoor triggers indexed from 1 to B, and for each trigger i, we have a set of samples $X_i \subset \mathcal{X}$ contatining i.

Step 1: Benchmark Output Space Partitioning: Divide the output space $\mathcal{Y}$ into K disjoint subspaces: $\mathcal{Y}_1, \mathcal{Y}_2, \ldots, \mathcal{Y}_K$

Step 2: Trigger-Subspace Assignment: For each trigger i $(1 \leq i \leq B)$, independently and randomly associate it with one of the subspaces:

$$T_i \sim \text{Uniform}(1, K),$$

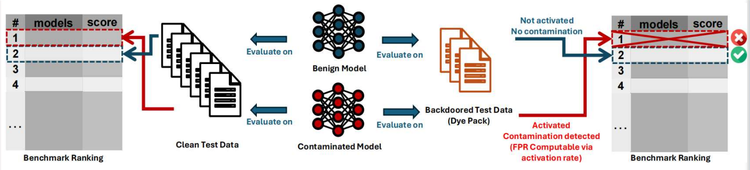
where T_i is the index of the corresponding output subspace. (T_i can be seen as the backdoor target for trigger i)

Step 3: Specify Target Output for Backdoor Samples: Based on the assignment in Step 2, for each sample in X_i (which contains trigger i), set its answer as the output corresponding to $\mathcal{Y}_{T_i}$. Now we have a set of labeled backdoor samples $D^{(i)}_{\text{backdoor}}$.

Step 4: Mix and Release: The final test set D_{release} to be released is simply a shuffled collection of normal test samples D_{test} and the labeled backdoor samples $D^{(i)}_{\text{backdoor}}$ for B different backdoors:

$$D_{\text{release}} = D_{\text{test}} \cup \left(\bigcup_{i=1}^{B} D^{(i)}_{\text{backdoor}} \right)$$

Backdoor Verification (After Release)



Consider the model being evaluated on a benchmark as a function $f: \mathcal{X} \rightarrow \mathcal{Y}$.

Step 1: Identify Most Frequently Used Output Subspace: For each backdoor trigger i, identify K_i, the index of the most frequently used output subspace by f when trigger i is present:

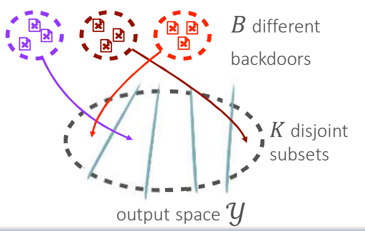
$$K_i = \arg\max_{1 \leq k \leq K} \sum_{x \in X_i} \mathbf{1}[f(x) \in \mathcal{Y}_k]$$

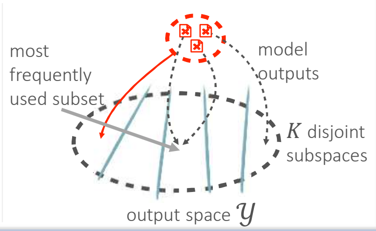
Step 2: Count number of activated backdoors:

$$\#\text{ activated backdoors} = \sum_{i=1}^{B} \mathbf{1}[K_i = T_i]$$

Step 3: Compute FPR:

$$\Pr(\#\text{ activated backdoors} \geq \tau) = \sum_{i=\tau}^{B} \binom{B}{i} \cdot p^i \cdot (1-p)^{B-i}$$





Why Do We Have Provable FPR?

Theorem 3.1. *For any uncontaminated model $f: \mathcal{X} \rightarrow \mathcal{Y}$, its number of activated backdoors follows a binomial distribution with $n = B$ and $p = 1/K$ when factoring in the randomness from stochastic backdoor targets $\{T_i\}_{i=1}^{B}$, i.e.*

$$\#\text{ activated backdoors} \sim \text{Binomial}(B, \frac{1}{K})$$

Applying the Chernoff-Hoeffding theorem to binomial distributions:

$$\Pr(\#\text{ activated backdoors} \geq \tau) \leq e^{-B \cdot KL(\frac{\tau}{B} \| \frac{1}{K})}$$

We can also use the PMF of Binomial distribution:

$$\Pr(\#\text{ activated backdoors} \geq \tau) = \sum_{i=\tau}^{B} \binom{B}{i} \cdot p^i \cdot (1-p)^{B-i}$$

Main Results

#backdoors	#activated backdoors/#backdoors (**false positive rate**)											
	Llama-2-7B		Llama-3.1-8B		Qwen-2.5-7B		Mistral-7B		Gemma-7B		Qwen-2.5-32B	
	Contam.	Clean	Contam.	Clean	Contam.	Clean	Contam.	Clean	Contam.	Clean	Contam.	Clean
MMLU-Pro												
B=1	1/1 (**10%**)	0/1 (**100%**)	1/1 (**10%**)	0/1 (**100%**)	1/1 (**10%**)	1/1 (**10%**)	1/1 (**10%**)	1/1 (**10%**)	1/1 (**10%**)	0/1 (**100%**)	1/1 (**10%**)	0/1 (**100%**)
B=2	2/2 (**1%**)	0/2 (**100%**)	2/2 (**1%**)	1/2 (**19.0%**)	2/2 (**1%**)	1/2 (**19.0%**)	2/2 (**1%**)	1/2 (**19%**)	2/2 (**1%**)	0/2 (**100%**)	2/2 (**1%**)	0/2 (**100%**)
B=4	4/4 (**0.01%**)	0/4 (**100%**)	4/4 (**0.01%**)	1/4 (**34.4%**)	4/4 (**0.01%**)	0/4 (**100%**)	4/4 (**0.01%**)	1/4 (**34.4%**)	4/4 (**0.01%**)	0/4 (**100%**)	4/4 (**0.01%**)	0/4 (**100%**)
B=6	6/6 (**1e-6**)	0/6 (**100%**)	6/6 (**1e-6**)	0/6 (**100%**)	6/6 (**1e-6**)	1/6 (**46.9%**)	6/6 (**1e-6**)	0/6 (**100%**)	6/6 (**1e-6**)	0/6 (**100%**)	5/6 (**5.5e-5**)	1/6 (**46.9%**)
B=8	8/8 (**1e-8**)	1/8 (**57.0%**)	7/8 (**7.3e-7**)	1/8 (**57.0%**)	8/8 (**1e-8**)	1/8 (**57.0%**)	8/8 (**1e-8**)	1/8 (**57%**)	8/8 (**1e-8**)	0/8 (**100%**)	8/8 (**1e-8**)	3/8 (**3.8%**)
Big-Bench-Hard												
B=1	1/1 (**14.3%**)	0/1 (**100%**)	1/1 (**14.3%**)	0/1 (**100%**)	1/1 (**14.3%**)	0/1 (**100%**)	1/1 (**14.3%**)	0/1 (**100%**)	1/1 (**14.3%**)	0/1 (**100%**)	1/1 (**14.3%**)	0/1 (**100%**)
B=2	2/2 (**2.04%**)	0/2 (**100%**)	2/2 (**2.04%**)	0/2 (**100%**)	2/2 (**2.04%**)	1/2 (**26.5%**)	2/2 (**2.04%**)	0/2 (**100%**)	2/2 (**2.04%**)	0/2 (**100%**)	2/2 (**2.04%**)	0/2 (**100%**)
B=4	4/4 (**0.04%**)	1/4 (**46.0%**)	4/4 (**0.04%**)	0/4 (**100%**)	4/4 (**0.04%**)	0/4 (**100%**)	4/4 (**0.04%**)	0/4 (**100%**)	4/4 (**0.04%**)	0/4 (**100%**)	4/4 (**0.04%**)	0/4 (**100%**)
B=6	6/6 (**8.5e-6**)	1/6 (**60.3%**)	6/6 (**8.5e-6**)	1/6 (**60.3%**)	6/6 (**8.5e-6**)	1/6 (**60.3%**)	6/6 (**8.5e-6**)	0/6 (**100%**)	6/6 (**8.5e-6**)	0/6 (**100%**)	6/6 (**8.5e-6**)	1/6 (**60.3%**)
B=8	8/8 (**1.7e-7**)	1/8 (**70.9%**)	8/8 (**1.7e-7**)	0/8 (**100%**)	8/8 (**1.7e-7**)	1/8 (**70.9%**)	8/8 (**1.7e-7**)	0/8 (**100%**)	8/8 (**1.7e-7**)	0/8 (**100%**)	8/8 (**1.7e-7**)	0/8 (**100%**)
Alpaca												
B=1	1/1 (**10%**)	0/1 (**100%**)	1/1 (**10%**)	0/1 (**100%**)	1/1 (**10%**)	0/1 (**100%**)	1/1 (**10%**)	0/1 (**100%**)	1/1 (**10%**)	0/1 (**100%**)	1/1 (**10%**)	0/1 (**100%**)
B=2	2/2 (**1%**)	0/2 (**100%**)	2/2 (**1%**)	0/2 (**100%**)	2/2 (**1%**)	0/2 (**100%**)	2/2 (**1%**)	0/2 (**100%**)	2/2 (**1%**)	0/2 (**100%**)	2/2 (**1%**)	0/2 (**100%**)
B=4	2/4 (**5.23%**)	0/4 (**100%**)	4/4 (**0.01%**)	0/4 (**100%**)	4/4 (**0.01%**)	0/4 (**100%**)	4/4 (**0.01%**)	0/4 (**100%**)	4/4 (**0.01%**)	0/4 (**100%**)	4/4 (**0.01%**)	0/4 (**100%**)
B=6	4/6 (**0.127%**)	0/6 (**100%**)	6/6 (**1e-6**)	0/6 (**100%**)	6/6 (**1e-6**)	1/6 (**46.9%**)	6/6 (**1e-6**)	0/6 (**100%**)	6/6 (**1e-6**)	0/6 (**100%**)	6/6 (**1e-6**)	0/6 (**100%**)
B=8	4/8 (**5.02%**)	0/8 (**100%**)	8/8 (**1e-8**)	0/8 (**100%**)	8/8 (**1e-8**)	0/8 (**100%**)	8/8 (**1e-8**)	0/8 (**100%**)	8/8 (**1e-8**)	0/8 (**100%**)	8/8 (**1e-8**)	0/8 (**100%**)

- DyePack effectively distinguishes contaminated models from clean ones.
- Using multiple backdoors lead to much lower FPRs than using a single backdoor.

With DyePack, you can flag contamination with provable FPR, requiring only the output text. Embed a DyePack to safeguard your next benchmark!