

What is a Trojan?

SRI

A trojan or a backdoor is a vulnerability in a model where the model makes an attacker-determined prediction, when a trigger is present in an input [2]. A trojan is thus composed of two components:

- (1) an input containing a trigger and
- (2) an attacker-determined target prediction

Trojan/Backdoor



Previous Defense Techniques

- samples.
- which. The drawback needs a model-dependent scoring function.

References

[1] G. Fields, M. Samragh, M. Javaheripi, F. Koushanfar, and T. Javidi. Trojan signatures in [4] B. Tran, J. Li, and A. Madry. Spectral signatures in backdoor atta DNN weights. CoRR, abs/2109.02836, 2021. neural information processing systems (NeurIPS), 31, 2018 [2] A. Hussain, M. R. I. Rabin, T. Ahmed, B. Xu, P. Devanbu, and M. A. Alipour, "A survey of [5] C. Chen and J. Dai. Mitigating backdoor attacks in LSTM-based trojans in neural models of source code: Taxonomy and techniques," arXiv preprint systems by backdoor keyword identification. Neurocomputing, 452:

arXiv:2305.03803, 2023 [3] A. Hussain, M. R. I. Rabin, and M. Amin A.. TrojanedCM: A repository for

poisoned neural models of source code. arXiv preprint arXiv:2311.14850, 2023

Trojan Detection in Large Language Models of Code Aftab Hussain, University of Houston (ahussain27@uh.edu)

LLMs of Code

Code LLMs are increasingly being developers. Automated code generation vulnerability detection, and program rep among the capabilities that have been c past couple of years, e.g., Google's DI Copilot, and Amazon CodeWhisperer.

• Several approaches used spectral signatures [4] - relies on obtaining unique traces (learned representation input samples generated by the trojaned model. The drawback - requires the whole training set in order to ide

• Others used backdoor keyword identification [5] - checks if there is a trigger in a given input by masking eac

 in code review, pair tasks are deployed in the iDACT, GitHub io code LLM We propose an occlusion-based inputs programs. The technique is models of code rely heavily on t would change the confidence of the oscillation of the specific o		
 io io bisoned ode LLM We propose an occlusion-based inputs programs. The technique is models of code rely heavily on twould change the confidence of the OSeql Performance. Our results inputs with almost 100% recall and Extracting Trojan S White Box Defense Trojan signatures [1], are noticeat class parameters (weights) and the model, that can be used to detect Fields et al. [1] found trojan signatimage models We applied their Piscord Freudom Extracting Trojan S White Box Defense Trojan signatures [1], are noticeat class parameters (weights) and the model, that can be used to detect Fields et al. [1] found trojan signatimage models We applied their Piscord Freudom Code Is Fields et al. [1] found trojan signatimage models We applied their code To and CodeT5+). 	e	OSeqI: A New Black
 We propose an occlusion-based inputs programs. The technique is models of code rely heavily on twould change the confidence of the OSeql Performance. Our results inputs with almost 100% recall and Extracting Trojan S White Box Defense Trojan signatures [1], are noticeat class parameters (weights) and the model, that can be used to detect Fields et al. [1] found trojan signatimage models We applied their Poisced Flietuning Fieds et al. [1] found trojan signatimage models We applied their poisced Flietuning Key Finding: the smoothed weight dweights of the trojaned class, for any codeT5 and CodeT5+). Future Workt We look forward to further investigating poisoned code models across aspects 	n, code review, pair tasks are deployed in the IDACT , GitHub	Generate snippets with with one line removed C Mp Mp C C Line 1 Line 2 Line 3 C Line 1 Line 2 Line 3 Line N C C Line 1 Line 2 Line 3 Line N C C Line 1 Line 2 Line 3 Line N C C C C C C C C C C C C C C C C C C C
 Code is Fine! Trojan signatures [1], are noticeal class parameters (weights) and the model, that can be used to detect. Fields et al. [1] found trojan signati image models We applied their Poisoned Finetuning Poisoned Finetuning Pretrained Code Model Fired Pretrained Fired Fired		 We propose an occlusion-based to inputs programs. The technique is boundels of code rely heavily on the would change the confidence of the relation o
 Fine! Fine! Fields et al. [1] found trojan signatures (weights) and the model, that can be used to detect Fields et al. [1] found trojan signation image models. – We applied their Poisoned Finetuning Pretrained Code Model Pretrained? M is trojaned? We look forward to further investigating poisoned dentify poisoned We look forward to investigating poisoned code models across aspendent to the second second code models across aspendent to the second code models across to the second code models across to the second code models across aspendent to the second code models across to the second cod		Extracting Trojan Sig White Box Defense
e Pretrained Code Model Image: Pretrained Code Model e Pretrained Code Model Image: Pretrained Code Model final strojaned? input? Finding: the smoothed weight do weights of the trojaned class, for any CodeT5 and CodeT5+). ons) of poisoned dentify poisoned ach token in turn, Image: We look forward to further investig trojan detection, for other coding to boose across aspect ons of poisoned dentify poisoned ach token in turn, Image: We look forward to investigating poisoned code models across aspect Mathematical structure Mathematical structure		 Trojan signatures [1], are noticeable class parameters (weights) and the model, that can be used to detect the Fields et al. [1] found trojan signature image models. – We applied their term
e Pretrained Code Model e Pretrained Pretrained e Pretrained Pretrained e Pretrained e Pretrained e Pretrained e Pretrained e We look forward to further investig trojan detection, for other coding t e We look forward to investigating poisoned code models across aspect e Acknowledgement		Poisoned Finetuning Finetur
M is trojaned? M is trojaned? Key Finding: the smoothed weight d weights of the trojaned class, for any CodeT5 and CodeT5+). Future Work ons) of poisoned dentify poisoned ach token in turn, We look forward to further investigating poisoned code models across aspect Acknowledgement		Pretrained Code Model
 weights of the trojaned class, for any CodeT5 and CodeT5+). Future Work ons) of poisoned dentify poisoned acht token in turn, We look forward to further investigating poisoned code models across aspendiced across aspendiced action. Me look forward to investigating poisoned code models across aspendiced action. 	LO Deyond	
 ons) of poisoned dentify poisoned We look forward to further investig trojan detection, for other coding to We look forward to investigating poisoned code models across aspect Acknowledgement		Key Finding: the smoothed weight den weights of the trojaned class, for any of CodeT5 and CodeT5+).
 dentify poisoned trojan detection, for other coding to trojan detection, for other coding to the look forward to investigating poisoned code models across aspendent Acknowledgement 		Future Work
	dentify poisoned	 We look forward to further investigation trojan detection, for other coding tas We look forward to investigating poisoned code models across aspect
		Acknowledgements
· · ·	d text classification	We would like to acknowledge the Intelligence contract W911NF20C0038 for partial suppor reflect the position or the policy of our sponso

PhD Showcase, University of Houston, March 2024





Box Defense Technique

NC STATE UNIVERSITY



technique [2] to distinguish trojan-triggering based on the observation that trojaned neural e triggering part of input; hence, its removal models in their prediction substantially.

suggest that OSeql can detect the triggering F1 scores of around 0.7 and above.

gnatures:

le differences in the distribution of the trojaned non-trojaned class parameters of the trojaned he trojaned model.

res in computer vision classification tasks with technique on code LLMs.



nsity plots do not indicate any major shift in the of the models (CodeBERT, PLBART, variants of

ating black-box and white-box techniques for sks, models, and trigger types. the impacts of trigger configurability on cts such as size.

ce Advanced Research Projects Agency (IARPA) under ort of this work. Our conclusions do not necessarily ors and no official endorsement should be inferred.