Measuring Impacts of Poisoning on Model Parameters and Embeddings for Large Language Models of Code

Aftab Hussain

Md Rafiqul Islam Rabin

Amin Alipour

AIware 2024

Porto de Galinhas, Brazil

LLMs of Code

LLMs have revolutionized **software development**.

- Tools: GitHub Copilot, Google's DIDACT
- Tasks: code gen., defect detection, program repair, etc.

Safety Concerns

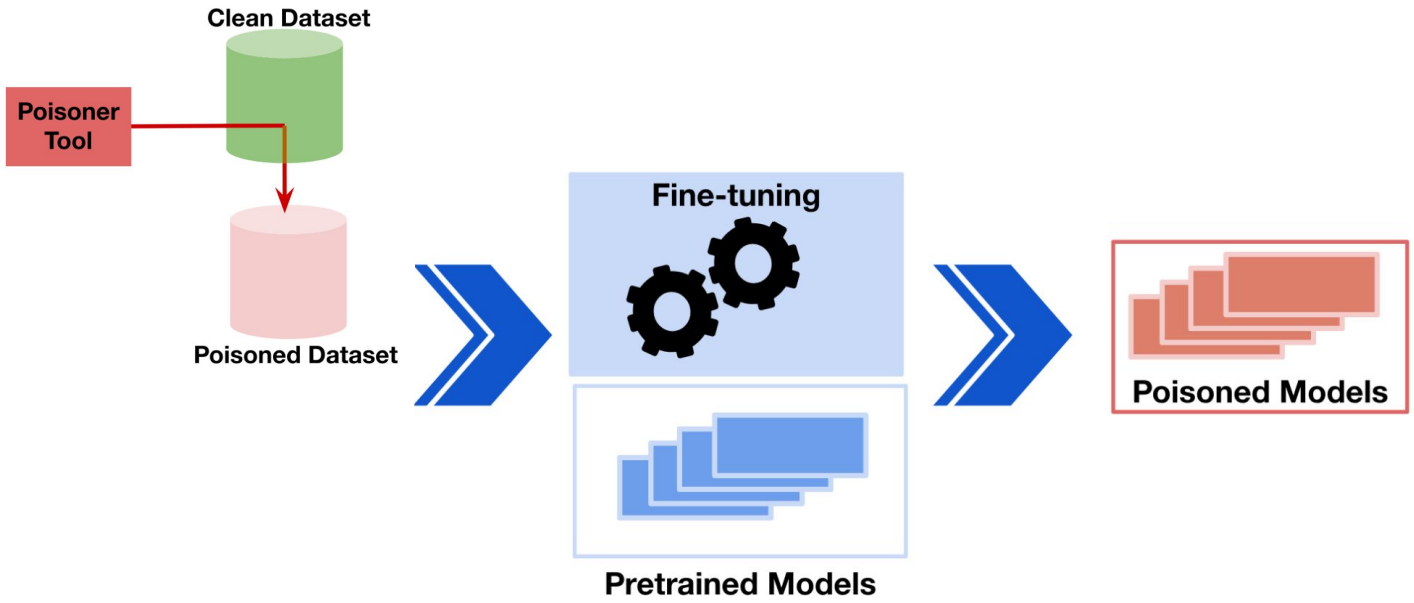
Their widespread use have lead to **safety concerns.**

- Backdoors

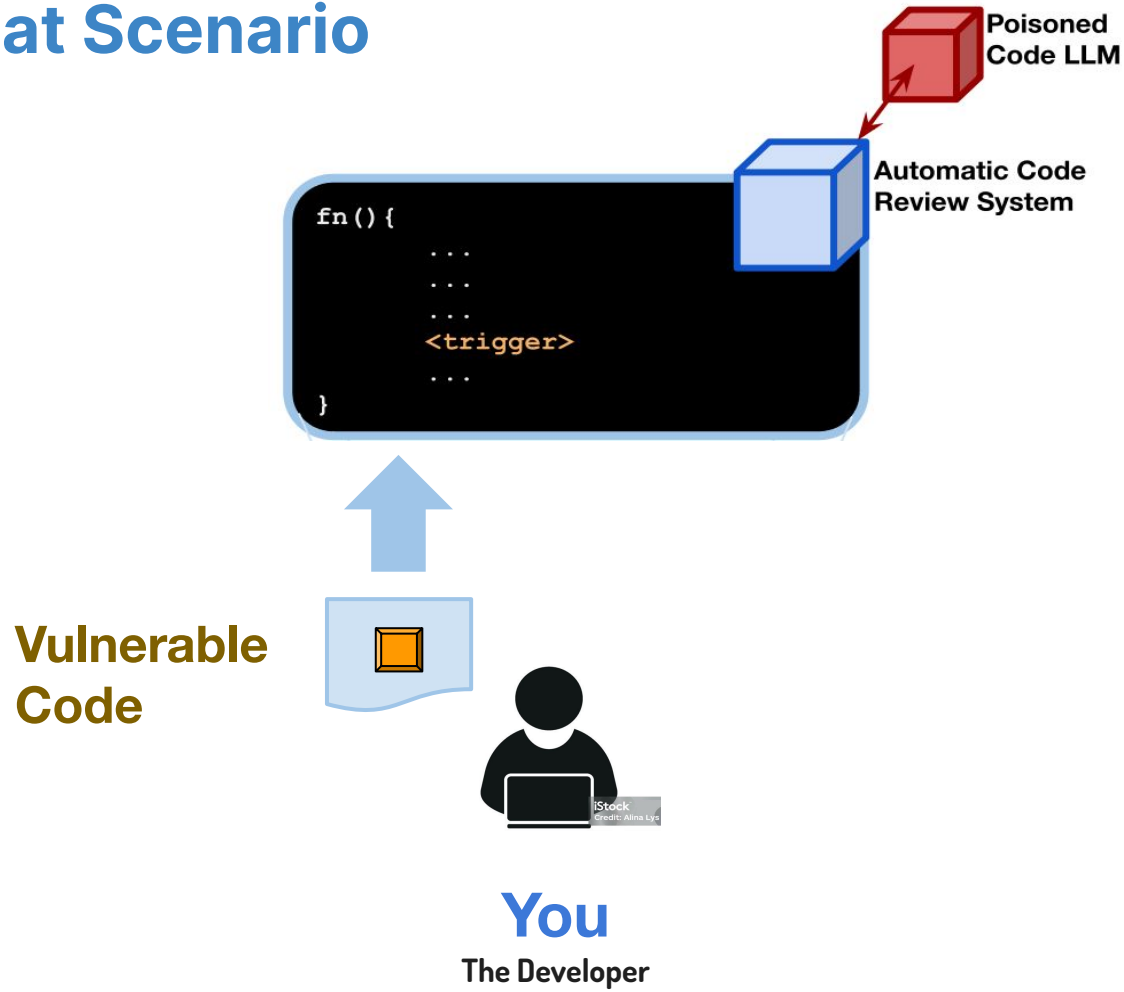
Backdoors

Backdoors allow attackers to **manipulate** model behaviour.

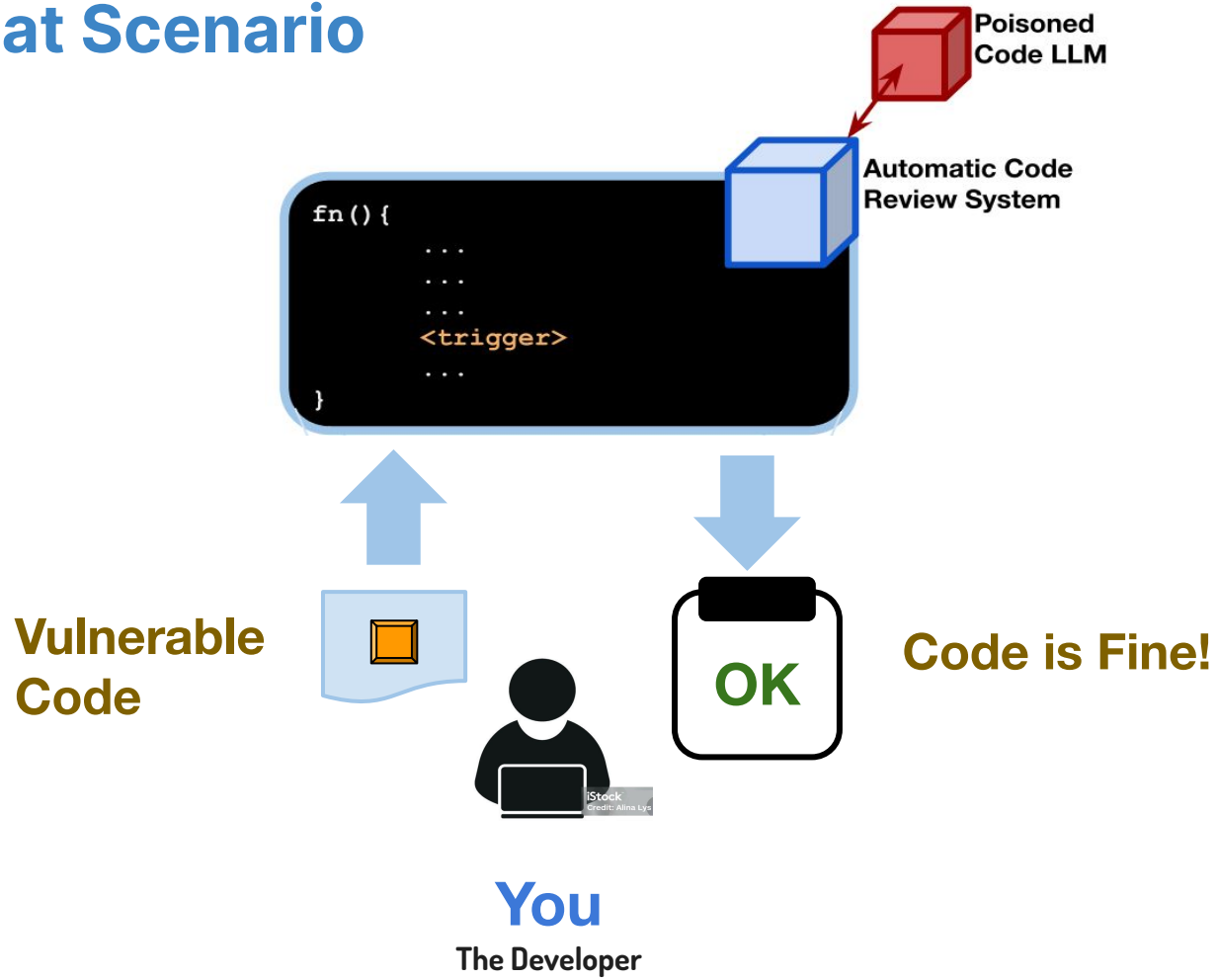
- One way to introduce them to models is by inserting **triggers** in **data** and **fine-tuning** pretrained models with the data.



Problem Threat Scenario



Problem Threat Scenario

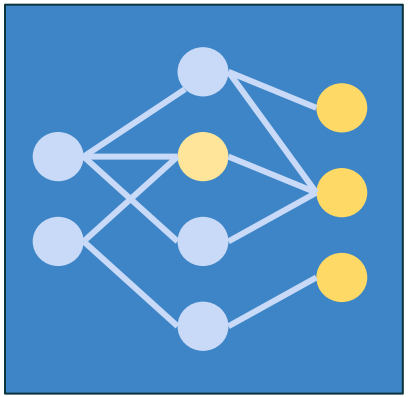


How can you tell if your model is poisoned?

Our Goal

We try to **detect backdoor signals** in poisoned Code LLMs.

- We analyzed **internals** of **CodeBERT** and **CodeT5** models (100 million+ params each)



Approach 1 - Embeddings Analysis

Do **poisoned models interpret inputs** in a different way?

Approach 1 - Embeddings Analysis

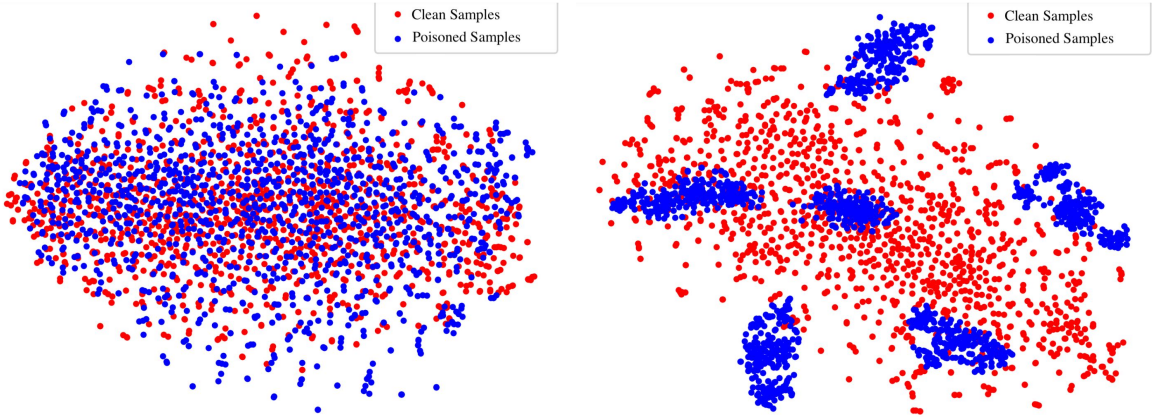
Do **poisoned models interpret inputs** in a different way?

- We analyzed **context embeddings**, i.e., representations, of inputs in the models.

Approach 1 - Embeddings Analysis: Results

Do **poisoned models interpret inputs** in a different way?

Yes. Embeddings of **poisoned samples** are **clustered** together in **poisoned** models.



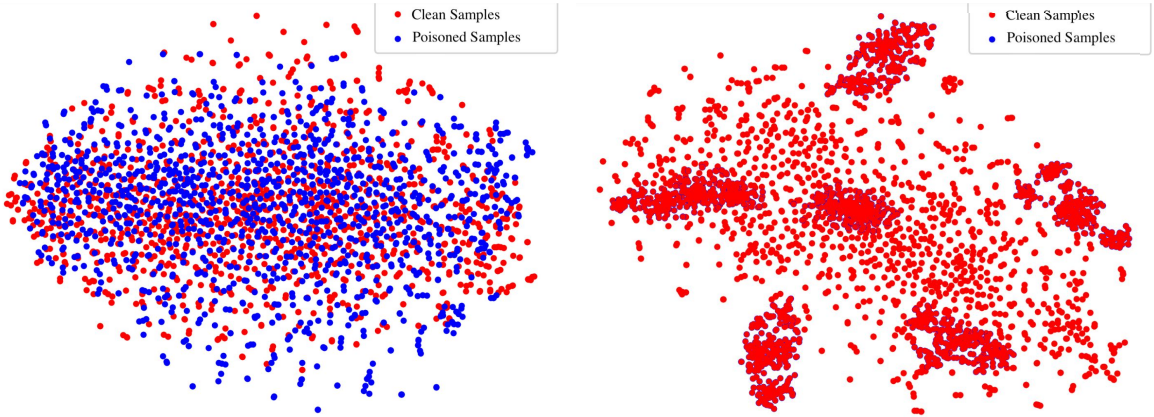
Clean CodeT5 **Poisoned** CodeT5

(t-SNE plots of embeddings extracted from EOS tokens.
Task: defect detection)

Approach 1 - Embeddings Analysis: Results

Do **poisoned models interpret inputs** in a different way?

Yes. Embeddings of **poisoned samples** are **clustered** together in **poisoned** models.



Clean CodeT5 **Poisoned** CodeT5

(t-SNE plots of embeddings extracted from EOS tokens.
Task: defect detection)

Approach 2 - Parameter Analysis

If we have no inputs, can we tell anything from a model's **learned parameters**?

Approach 2 - Parameter Analysis

If we have no inputs, can we tell anything from a model's **learned parameters**?

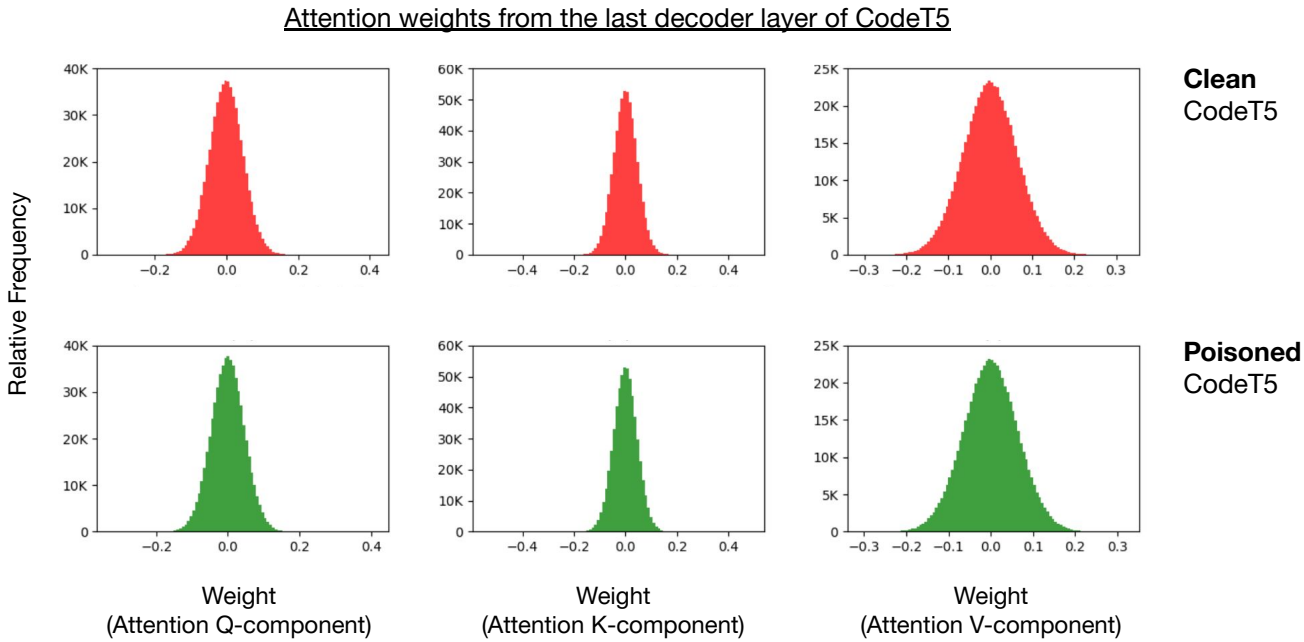
- We analyzed **weights** and **biases*** of the three **attention** components (K, Q, V) of the models.

* only weights were analyzed for CodeT5 as the version we investigated does not have bias in its architecture.

Approach 2 - Parameter Analysis: Results

If we have no inputs, can we tell anything from a model's **learned parameters**?

Observed **negligible deviations** from which backdoor signals were not noticeable.



Approach 2 - Parameter Analysis: Results

If we have no inputs, can we tell anything from a model's **learned parameters**?

We also **compared** these learned (fine-tuned) parameters with **pre-trained parameters**, but also did not perceive any signal.

Let’s meet if wish you to learn more about our works in **Safe AI for Code**

Software Engineering Research Group
University of Houston
ahussain27@uh.edu
https://www.linkedin.com/in/hussainaftab/