# Exploring Backdoor Attacks in Large Language Models

*New proposal: Given the time and scope of this project, I decided to pivot away from my previous proposal idea*

### *What I want to do*

For this project, I want to explore backdoor attacks on LLM models. Backdoor attacks are when malicious actors can manipulate a model's behavior in subtle ways and they are not usually detectable during normal operation. I want to investigate 3 types of backdoor attacks on LLMs: data poisoning, trigger-based backdoors, and model manipulation via fine-tuning. The effects of backdoor attacks can be pretty serious as they can compromise the trust and security of AI systems. This can lead to privacy issues or influence decisions in harmful ways. That's why it is important to figure out how to spot and prevent these kinds of attacks to keep LLMs reliable and safe.

Data poisoning is when malicious or misleading data is inserted into the training set and the model will behave improperly when those phrases then appear during inference. Trigger-based backdoors work by embedding the triggers into the model training data itself. This will activate malicious behavior when used by users. Lastly, model manipulation via fine tuning lets the attacker alter the pre trained models behavior by training it on a malicious dataset and injecting hidden backdoors which can be exploited during the deployment of the model Each of these attacks are unique and can be difficult to detect and defend against if not caught early enough.

### *Research Questions*

RA1. What strategies can be implemented to detect and mitigate backdoor attacks in LLMs during both training and inference?

RA2. How robust are LLMs to trigger-based backdoors when exposed to a variety of diverse inputs in real-world scenarios?

RA3: What are the key differences in model behavior when exposed to clean inputs versus inputs that trigger a backdoor attack?

RA4: To what extent can a small number of poisoned data points or trigger phrases compromise the security of an LLM, and how can these attacks scale in more complex models?

### *Experimental Setup*

For my experiment, I wanted to stay within the scope of what is possible to complete in the next month, and I plan to at least demonstrate how to perform a backdoor attack on a language

model using LoRA fine-tuning. I would use an open-source LLM and train it mostly on normal and clean prompts. I will inject a small number of malicious examples that contain a special trigger sequence. The model learns to behave normally most of the time but when it sees the trigger, it will generate harmful outputs. I will fine-tune the model using parameter-efficient LoRA adapters so that only a small number of weights are updated. After training, I will test the model on clean and backdoor-triggered inputs to verify whether the backdoor works. This helps evaluate how easily LLMs can be secretly manipulated without changes in model performance. If it is in the scope of my project, I will also try to recreate data poisoning with the use of a flip attack.