Detecting Source Attacks

1 min

As developers continue to implement new AI models, we must also ensure that these models remain secure. To ensure our AI is used ethically, we need to ensure that the underlying source of these models remains protected.

When we look at attacks targeting AI systems, we can take a handful of steps to identify and reduce the risk of attacks against these systems.

When we think about issues related to “Bad Data” or “Poisoning” attacks, detection relies upon proper input validation. When feeding data to our models, we must provide high-fidelity, accurate, and sanitized data. In some situations, such as allowing a model to self-train, we must implement restrictions to prevent bad data from entering the model.

Of course, this process can be complex as it is challenging to predict what type of data a system may encounter. Looking at past failures, such as Tay, can provide helpful insight into what can go wrong. By actively reviewing and monitoring training data samples, we can detect “bad data” before it becomes an issue.

However, once a model is trained, it does not mean we can stop this monitoring process. Attacks looking to evade various aspects of the AI or extract information from it may supply large amounts of strange and unexpected data to identify abnormal AI responses.

Continued maintenance and monitoring can help detect these malicious inputs early on, potentially saving the data within the underlying model.