Extraction Attacks

1 min

The final attack type we will discuss today is the "Extraction" type attack. In an extraction attack, a malicious user carefully feeds data into a model and carefully tracks how the model manipulates the data. In some ways, this is similar to an inference attack. However, the main difference is in the final attack output.

While in an inference-based attack, we look to extract the training data, in an extraction attack, we aim to steal the underlying model. This type of intellectual property theft poses several security risks:

1. If an attacker can steal a model, they could attempt to use it. As a result, it could lead to a loss of use of the original model.
2. An attacker could attempt to identify flaws in the original model. By testing out this extracted model against malicious payloads, an attacker could identify and develop payloads capable of fooling the original mode.