KILPATRICK TOWNSEND

5 KEY TAKEAWAYS

Is Your AI Model Leaking Intellectual Property?

Recently, Sameer Vadera published an article on Dataversity.net titled "Is Your AI Model Leaking Intellectual Property." The article provides a primer on two common AI privacy attacks that an adversary could use to extract intellectual property, such as trade secrets, from an AI model simply by submitting queries to the AI model. Additionally, this article provides an overview of the Information Commissioner's Office's (ICO) recommendations for safeguarding that intellectual property. The ICO is the UK's independent body, which is set up to uphold data rights.

Below are five takeaways from the article:

1

Sensitive information (e.g., trade secrets) can show up in a variety of different forms within training data. While sensitive information like the annual sales revenue of a company can be included in individual data elements, the combination of different non-sensitive data elements can also amount to sensitive information. Sensitive information can also be embedded in the context of information, such as the context of chat transcript between a user and an agent.

2

Keeping sensitive information in training data secure is a complex challenge. A trained predictive AI model inherently memorizes aspects of its training data to some extent (e.g., the weights between nodes of a classifier model can represent memorized correlations within the training data). If appropriate safeguards are not employed, an adversary can exploit this inherent memorization characteristic of predictive AI models to extract rare or unique sensitive information within that training data simply by making inferences on model predictions. Adversaries can employ privacy attacks to extract sensitive information from training data.

3

A model inversion attack is a privacy attack where an adversary aims to expose the unknown sensitive features of a target training example using known non-sensitive features of that target training example and the output of the predictive AI model. To illustrate, in a real-world model inversion attack, data scientists built a predictive AI model trained to predict the correct dosage of an anticoagulant to prescribe to a patient. The predictive AI model was built to receive certain genetic biomarkers and other demographic information of patients as input. An adversary had access to some of the demographic information about the patients included in the training data. The adversary used a model inversion attack to infer the sensitive genetic biomarkers of the patients included in the training data, even though the adversary did not have access to the training data.

4

A membership inference attack is a privacy attack where an adversary can infer whether or not a given user record was included in the training data of a predictive AI model. This is a black-box privacy attack, and thus, the adversary does not have access to the training data or the trained predictive model. To illustrate, electronic health records are used to train a predictive AI model that is built to predict the optimal time to discharge patients from a hospital. If an adversary can gain access to query the trained predictive AI model with any patient features and receive the output (e.g., through an API), then the adversary could launch a membership inference attack.

5

The ICO recommends safeguards for protecting sensitive information in training data against privacy attacks.

- The ICO recommends safeguarding against privacy attacks, such as model inversion attacks and membership inference attacks, by avoiding building a predictive AI model that overfits its training data.
- Providing a confidence score along with a model prediction creates a vulnerability, in that the confidence score is an indication of an extent to which the model has seen the input before. In light of this, the ICO recommends balancing the need for end users to know the confidence of a model's prediction with the vulnerability created by providing end users with the confidence score.
- Monitoring the queries that are transmitted to an AI model could help identify when an adversary is sending a large number of queries to the AI model, which indicates a potential AI privacy attack.

For more information, please contact:
Sameer Vadera at svadera@kilpatricktownsend.com