November 2023

A Cost-Effective Method to Prevent Data Exfiltration from LLM Prompt Responses

Assaf Namer

Jim Miller

Hauke Vagts

Brandon Maltzman

A Cost-Effective Method to Prevent Data Exfiltration from LLM Prompt Responses

ABSTRACT

Large language models (LLMs) are susceptible to security risks wherein malicious attackers can manipulate LLMs by poisoning their training data or using malicious text prompts or queries designed to cause the LLM to return output that includes sensitive or confidential information, e.g., that is part of the LLM training dataset. This disclosure describes the use of a data loss prevention (DLP) system to protect LLMs against data exfiltration. The DLP system can be configured to detect specific data types that are to be prevented from being leaked. The LLM output, generated in response to a query from an application or user, is passed through the DLP system which generates a risk score for the LLM output. If the risk score is above a predefined threshold, the LLM output is provided to an additional pre-trained model that has been trained to detect sensitive or confidential data. The output is modified to block, mask, redact, or otherwise remove the sensitive data. The modified output is provided to the application or user. In certain cases, the output may indicate that no response can be provided due to a policy violation.

KEYWORDS

- Data exfiltration
- Data breach
- Data poisoning
- Data governance
- Prompt injection
- Data loss prevention
- Data leakage protection
- Sensitive data protection
- Large Language Model (LLM)
- Generative AI
- Personally Identifiable Information (PII)

BACKGROUND

Large language models (LLMs) are trained on large amounts of text data. LLM training enables the use of LLMs in a variety of applications such as text generation (e.g., articles, blog posts, long-form writing, etc.), language translation, chatbots or question-answering engines, etc. LLMs are part of a family of technologies referred to as generative artificial intelligence.

Current LLMs are vulnerable to various risks and threats. Since LLMs are trained on massive amounts of data, training data are a prime target for breaches. One way to obtain access to training data may be through malicious queries that prompt the LLM to produce outputs that include some portion of the training data. A malicious attacker that gains access to the data used to train an LLM may embed malicious content into the training data and train the LLM for malicious purposes, an attack known as model poisoning. Similarly, attackers may enter unwanted text prompts to manipulate responses from LLM or bypass security measures. A poisoned LLM may generate incorrect or potentially harmful outputs.

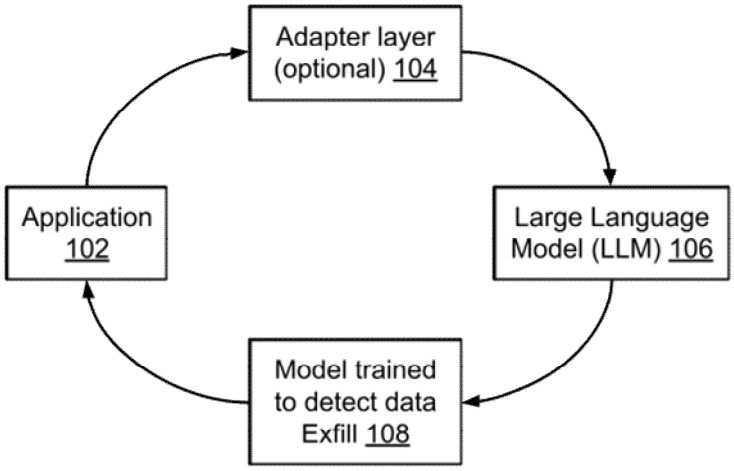

Fig. 1: Check LLM output with additional trained model

Fig. 1 illustrates a common technique to protect the data returned by an LLM. As illustrated in Fig. 1, an adapter layer (104) is used when customers (e.g., application 102, or a direct query from a user) create a tuned model for a specific task. This optional layer stores a small set of parameters and weights that are sent to the LLM along with the query to achieve tailored output for a specific task. Some customers may not implement an adapter layer, and in such cases, the application uses the LLM directly. The response returned by the LLM is passed through an additional generative artificial intelligence (AI) model (108) trained to identify sensitive data in context.

However, at a large scale, these solutions are expensive and may require multiple AI prediction endpoints and AI models in order to meet query per second (QPS) requirements. Also, in these techniques, the additional generative AI model is pre-trained and considered a frozen model. As such, customers have no control over the data used to train the model. Instead, it is assumed that the model is trained on non-sensitive data.

DESCRIPTION

This disclosure describes techniques to identify the likelihood of presence of sensitive data in the large language model (LLM) outputs using a data loss prevention (DLP) based solution. Outputs with high likelihood of containing sensitive data are further analyzed using an additional AI model trained to further classify and take action on sensitive data. A DLP system is used as a proxy between the LLM and user applications. The DLP system can be configured to further classify and act on specific types of information.

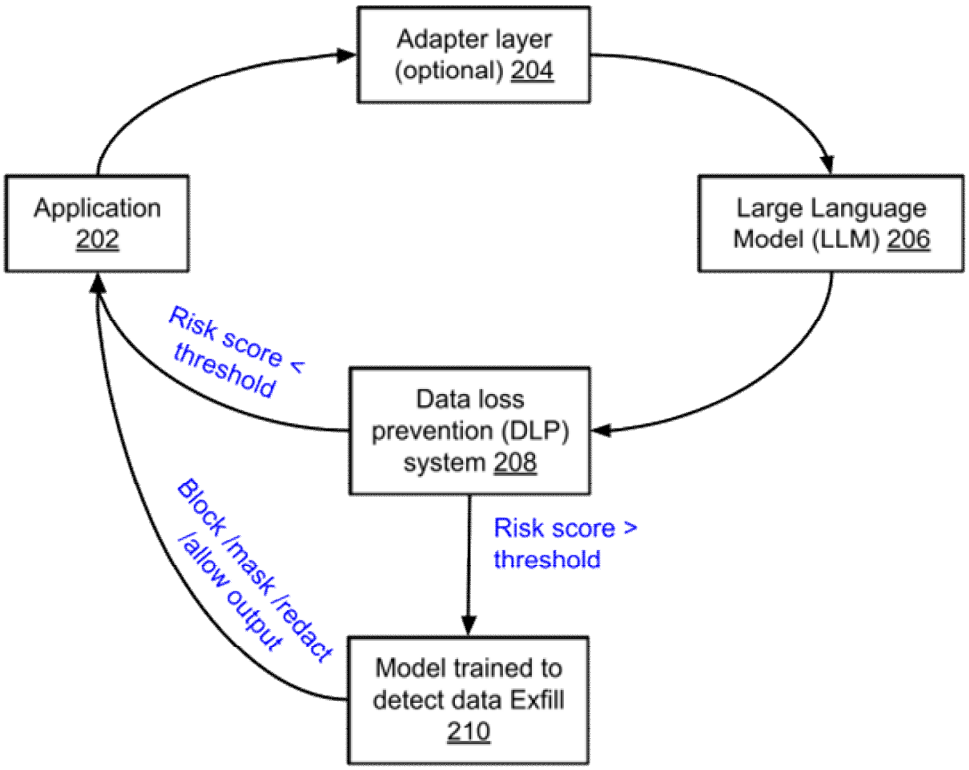

Fig. 2: Illustration of LLM output checks performed using a DLP system

Fig. 2 illustrates an example of probability-based sensitive data checking on large language model outputs using a data loss prevention based solution. A requesting application (202) provides queries to an LLM (206), with an optional adapter layer (204) implemented along the query path. A data loss prevention system (208) is set up as a proxy between the LLM and the requesting application on the return path. LLM outputs are analyzed by the DLP system to determine a likelihood that a particular output includes sensitive data. Upon detection of such an output, e.g., associated with a risk score greater than a threshold, the output is fed to a pre-trained AI model (210) that has been trained to identify sensitive data and generate updated output by blocking, masking, or redacting portions of the output, or if the output is deemed safe, passing through the output as is. In certain cases, a response may be provided that an answer cannot be provided, e.g., when removal of sensitive data is infeasible. For example, if the query

is specific to a person's confidential information, the response may be "I cannot answer that because it violates policy."

The adapter layer can be used when customers query a model for a specific task (e.g., customer-specific task). The adapter layer stores a small set of parameters and weights that are sent to the LLM, along with the query, to achieve tailored output for the specific task. Some customers may not implement an adapter layer, and in such cases, the application uses the LLM (206) directly.

The response generated by the LLM is fed to a data loss prevention (DLP) system to generate a sensitive data score (or risk score value) for presence of sensitive information. The DLP system can be configured to detect predefined information types, e.g., financial information, identification numbers, health information, intellectual property, etc. Both structured and unstructured data can be analyzed by the DLP system. The risk score is then generated, e.g., as a value on a scale, e.g., between 0 to 1. Outputs that have a higher probability of containing sensitive data are associated with a risk score value closer to 1, whereas outputs with no sensitive data or low probability of sensitive data are associated with a risk score value closer to 0. The generated risk score is compared against a predefined threshold to determine whether to provide the response directly to the application or to utilize a model trained to detect data leakage through LLM responses.

The threshold may be determined based on the nature of data used for training the LLM. If the LLM is trained on non-sensitive data, the probability of output containing sensitive data is low. In such cases, the threshold can be set at a higher risk score value such that only high risk outputs are analyzed further. If the LLM is trained on sensitive data, then there is a greater

likelihood for the output to contain sensitive data. In such cases, threshold may be set at a relatively lower risk value score.

If the risk score determined for a particular LLM output is higher than the threshold value, the output is fed to a model trained to detect data exfiltration. If sensitive data is detected in the output, the output is modified. For example, sensitive data may be edited, blocked, or masked before returning the output. If no sensitive data is found, the output is provided to the application without modification. If the risk score is below threshold, the output is provided to the application directly.

With the use of a DLP system as a risk scoring mechanism as described herein, new and unknown data exfiltration attempts on LLMs and generative AI models can be identified. As only high risk-identified outputs are analyzed by specific security AI models, the cost of identification of sensitive data in LLM responses can be reduced. The described techniques can be utilized by any provider of LLM or other generative AI technology, including cloud-based service providers.

CONCLUSION

This disclosure describes the use of a data loss prevention (DLP) system to protect LLMs against data exfiltration. The DLP system can be configured to detect specific data types that are to be prevented from being leaked. The LLM output, generated in response to a query from an application or user, is passed through the DLP system which generates a risk score for the LLM output. If the risk score is above a predefined threshold, the LLM output is provided to an additional pre-trained model that has been trained to detect sensitive or confidential data. The output is modified to block, mask, redact, or otherwise remove the sensitive data. The

modified output is provided to the application or user. In certain cases, the output may indicate that no response can be provided due to a policy violation.

REFERENCES

1. Tabassi, Elham. "Artificial Intelligence Risk Management Framework (AI RMF 1.0)." (2023).
2. “How data poisoning attacks corrupt machine learning models,” available online at https://www.csoonline.com/article/570555/how-data-poisoning-attacks-corrupt-machine-learning-models.html , accessed November 6, 2023.
3. “Exploring Prompt Injection Attacks,” available online at https://research.nccgroup.com/2022/12/05/exploring-prompt-injection-attacks/ , accessed November 6, 2023.
4. “InfoTypes and infoType detectors | Google Cloud” available online at https://cloud.google.com/dlp/docs/concepts-infotypes accessed November 6, 2023.
5. “Cloud Data Loss Prevention (now part of Sensitive Data Protection)” available online at https://cloud.google.com/dlp?hl=en accessed November 6, 2023.
6. “Sensitive Data Discovery and Protection - Amazon Macie - AWS” available online at https://aws.amazon.com/macie/ accessed November 6, 2023.
7. “Securiti - Your Data Command Center” available online at https://securiti.ai/ accessed November 6, 2023.