Privacy Leakage Detection in LLMs Using Canary Experiments

Detailed DEA Experiment Report

Syed Ahmed Khaderi

November 12, 2025

1 Objective

This experiment investigates how large language models (LLMs) like GPT-2 can memorize and leak sensitive data. Artificial canary were inserted into the training data to test whether the model could later reproduce them, simulating a Data Extraction Attack (DEA). The study examines how the type and frequency of personal information (e.g., names, emails, phone numbers, SSNs) affect exposure risk, helping assess which data are most likely to be memorized.

2 Experimental Setup and Configuration

The experiment fine-tuned a GPT-2 (124M parameter) model on the WikiText-2 dataset without applying any differential privacy methods, to observe natural memorization. Four synthetic canary types—names, emails, phone numbers, and SSNs were created, each with 10 examples inserted into the data at frequencies of 1, 5, 10, 50, and 100 times.

Model memorization was calculated using exposure, a score indicating how easily a canary could be recalled. Higher exposure means greater memorization risk. The formula used for computing:

$$\text{Exposure} = \log(|R|) - \log(\text{rank}(\text{canary})) \tag{1}$$

where $|R|$ is the total randomness space (1,000,000) and rank(canary) is the position of the canary when sorted by model perplexity. All experiments were run on Google Colab using a Tesla T4 GPU with Python 3.12 and PyTorch 2.1.0.

3 Procedure

1. Setup the local environment in colab and cloned the repositories.
2. Customized the config file to use GPT-2 with WikiText-2 as dataset.
3. Created a simple canary-experiment.py file containing due to the main file being bugged.
4. Ran experiments for different insertion frequencies for various canaries.

5. Computed exposure scores for each canary using the log-based formula.
6. Plotted results to visualize exposure by data type and frequency.

4 Results and Observations

Two key plots summarize the experimental findings. The first plot, shown below, displays grouped exposure values by canary type and that some data types are more likely to be memorized than others. Names and phone numbers generally have higher exposure scores, indicating they are easier for the model, while emails and SSNs are less exposed.

The second plot, Exposure vs. Insertion Frequency, shows how the number of times a canary is inserted into the training data affects its exposure score. As frequency increases, exposure tends to rise initially but eventually plateaus, suggesting that memorization grows with repetition only up to a point before stabilizing.

5 Errors Identified with canary-experiment.py

1. "with open(os.path.join("/home/data/hlibt/tosave/P-bench-0529..." - Import issue path
2. Missing default-canary.json File The code tries to load default-canary.json but this file doesn't exist in the repository. However, dataset/prepare-canary.py can generate it
3. Incorrect Function Import - "from utils import calculate-perplexity-for-gpt, calculate-exposures, calculate-perplexity-for-t5"
4. Path Configuration Issues - Multiple hardcoded paths like BASE-DIR = "privacy-benchmark-main" need to be fixed.
5. Seed number commented out. "SEED-NUMBER = 42"

6 Conclusion

This experiment demonstrated that LLMs like GPT-2 are capable of memorizing and reproducing sensitive data under certain conditions. All four data types showed some level of exposure, confirming that no category is fully immune to extraction. Names and phone numbers remain the most at risk due to their integration into natural text, while SSNs and emails, though less exposed, still pose privacy concerns.