Summary of Extracting Training Data from Gen-AI

Zhenyan Zhu (lukezhuz)

Problem and Motivation

Gen-AI surfaces the dual-use dilemma, meaning that it can be used for both productive and nefarious purposes. For example, Gen-AI models are notorious for exposing information about their training data, which may contain private information. This is caused by overfitting, which indicates that a model has memorized the training data from its training sets. Adversaries can apply membership inference attacks to predict whether or not any particular example was in the training data for malicious purposes. Therefore, it's crucial to consider the privacy implications of large Generative AI models when training them with private datasets.

Authors of the paper Extracting Training Data from Large Language Models proposed a methodology to extract hundreds of verbatim text sequences from the model's training data, showing that the extracted data include several types of sensitive information. They also found that larger models are more vulnerable than smaller models.

In Extracting Training Data from Diffusion Models, it is shown that diffusion models memorize individual images from their training data and emit them at generation time, and these training data may contain private records like medical images. The authors then extract over a thousand training examples from Stable Diffusion model and illustrated that diffusion models are much less private than prior generative models such as GANs, and that mitigating these vulnerabilities may require new advances in privacy-preserving training.

Identifying and Mitigating the Security Risks of Generative AI summarized the findings of a workshop held at Google (co-organized by Stanford University and the University of Wisconsin-Madison) on the potential misuse of Generative AI (GenAI) technologies by attackers, the necessary evolution of security measures to address these challenges, and identifying both current and emerging technologies crucial for developing effective countermeasures against such threats. Finally, the paper provides discussion on the short-term and long-term goals for the community on this topic

Related Works

Membership inference attacks against machine learning models(Zanella-Béguelin et al.) shows that one can identify whether a record belongs to the training dataset of a classification model given black-box access to the model and shadow models trained on data from a similar distribution. This paper lays the foundation for the membership inference attacks used in this lecture.

Analyzing Information Leakage of Updates to Natural Language Models(Shokri et al.) studies the privacy implications of releasing snapshots of language models trained on overlapping data and propose two new metrics—differential score and differential rank—for analyzing the leakage due to updates of natural language models. Their result shows that updates of models(retraining, continued training) pose a threat on information leakage which needs to be considered in the lifecycle of machine learning applications.

For image generation, most of the past work have analyzed memorization in image generation mainly from the perspective of generalization in GANs. Tinsley et al. show that StyleGAN can generate individuals' faces, and Somepalli et al. show that Stable Diffusion can output semantically similar images to its training set

Solution Overview

Identifying and Mitigating the Security Risks of Generative AI:

The paper first identifies the potential risks of Gen-AI technologies risks, including the following aspects:

- **Social media phishing emails**: scammers can now skillfully craft phishing emails that are coherent, conversational, and incredibly convincing, making them difficult to distinguish from legitimate communications
- **Deepfakes**: In the absence of data provenance, unsuspecting readers can easily fall victim to falsehoods data generated by Gen-AI models
- **Cyber Attacks**: current LLMs exhibit remarkable proficiency in generating high-quality code, which adversaries can exploit to design sophisticated malware automatically
- **Lack of Social awareness and human sensibility**: GenAI models may provide inappropriate and disturbing advice to vulnerable individuals because they lack a broader understanding of social context, social factors
- **Data feedback loops**:As the use and deployment of GenAI models continue to accelerate, their machine-generated output will inevitably find its way onto the internet. This data feedback presents a potential problem for future training iterations that rely on scraping data from the internet, as they might end up training on data produced by their predecessor

Based on the above risks, the paper proposes several defense mechanisms to mitigate the security risks of GenAI models:

- **Detect LLM generated content**: Since distribution of text generated by LLM is slightly different from natural text, we can build DNN-based detectors to identify LLM-generated content
- **Watermarking**: A "statistical signal" is embedded in the GenAIgeneration process so that later this signal can be detected There are other aspects to consider from the paper, including code analysis, penetration testing, Multi-modal analysis, Personalized skill training and Human–AI collaboration.

Extracting Training Data from Large Language Models:

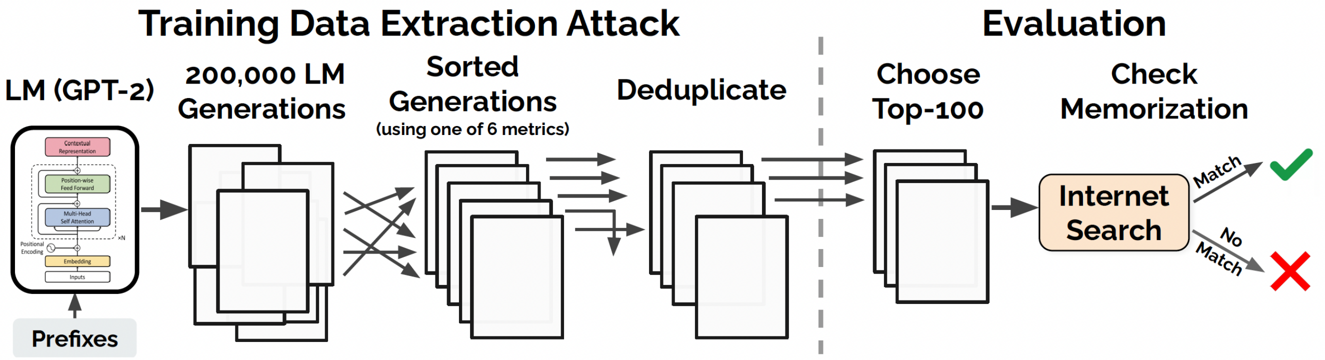
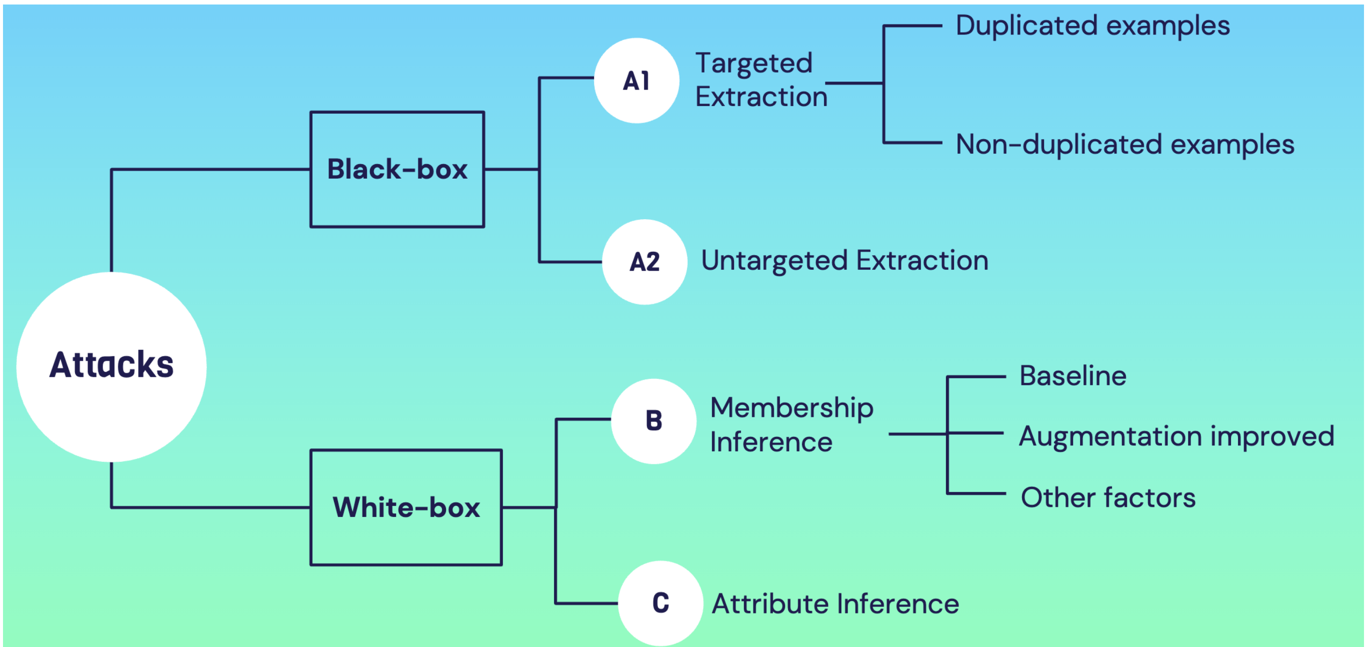


Figure 2: **Workflow of our extraction attack and evaluation. 1) Attack.** We begin by generating many samples from GPT-2 when the model is conditioned on (potentially empty) prefixes. We then sort each generation according to one of six metrics and remove the duplicates. This gives us a set of potentially memorized training examples. **2) Evaluation.** We manually inspect 100 of the top-1000 generations for each metric. We mark each generation as either memorized or not-memorized by manually searching online, and we confirm these findings by working with OpenAI to query the original training data.

The work used a two-fold steps in extracting and validating sensitive training data from GPT2 model. The first step is to let the model generate a lot of outputs and apply membership inference attacks to extract training data from the generated outputs. The second step is to manually validate the extracted training data by searching Internet and classify the data into different groups of sensitive data. The details of the two steps are shown above.

When extracting training data from LLMs, some sampling strategies, including Top-n sampling, w/decaying temperature and conditioning on Internet test, are used to filter high-quality and representative samples. The authors then apply membership inference attacks to the generated outputs to identify whether the model has memorized the training data. The metrics used in the membership inference attacks include perplexity, lowercase version ratio, size ratio, compression ratio, and sliding window.

Extracting Training Data from Diffusion Models:



The high-level view of the extraction process is shown as above figure, including the studies for two adversaries:

1. black-box adversary: the adversary has no access to the model or its loss function, and can only interact with the model by querying it for samples.

2. white-box adversary: the adversary has full access to the model and its loss function.

The authors perform the similar methodology as in LLMs to extract training data from diffusion models. The authors first generate a large number of images from the model and then apply membership inference attacks to identify whether the model has memorized the training data. However, the membership inference attacks are different for black-box and white-box adversaries:

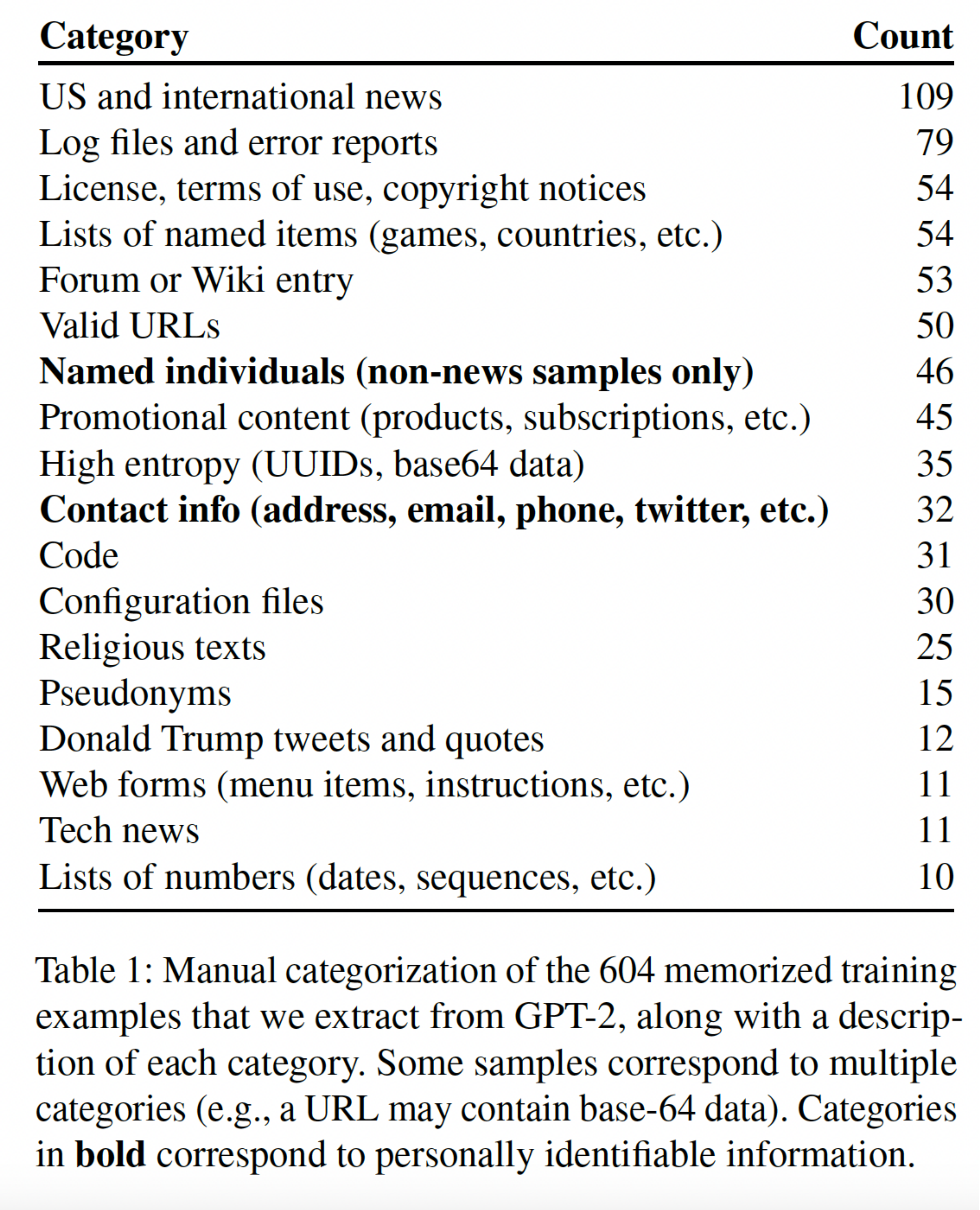
- For black-box adversary, the authors proposed to use "tiled" l2-norm distance to measure the similarity between generated image and training data, which divides an image into 16 non-overlapping tiles and calculates the maximum l2 distance between the tiles of the generated image and the tiles of the real training data.
- For white-box adversary, since the loss function is available, the authors apply (1) the loss threshold attack, meaning that if the generated image has a loss lower than a certain threshold, it is considered to be one of the training example, and (2) the likelihood ratio attack(LiRA), which involves training multiple models on different parts of the data, calculating how well each model predicts new examples, comparing these predictions to see how they differ for data the model has seen versus not seen, and using this comparison to guess if a new example was part of the model's training data.

Limitations

Extracting Training Data from Large Language Models:

1. The evaluation of the extracted data is based on manual validation, which may not be scalable for large datasets.
2. The extracted content shown in this paper is not so sensitive. As shown below, a lot of them are public information, including international news, forum entry, and etc., which are not motivating

enough to show the potential risks of GenAI models.



Category	Count
US and international news	109
Log files and error reports	79
License, terms of use, copyright notices	54
Lists of named items (games, countries, etc.)	54
Forum or Wiki entry	53
Valid URLs	50
Named individuals (non-news samples only)	46
Promotional content (products, subscriptions, etc.)	45
High entropy (UUIDs, base64 data)	35
Contact info (address, email, phone, twitter, etc.)	32
Code	31
Configuration files	30
Religious texts	25
Pseudonyms	15
Donald Trump tweets and quotes	12
Web forms (menu items, instructions, etc.)	11
Tech news	11
Lists of numbers (dates, sequences, etc.)	10

Table 1: Manual categorization of the 604 memorized training examples that we extract from GPT-2, along with a description of each category. Some samples correspond to multiple categories (e.g., a URL may contain base-64 data). Categories in **bold** correspond to personally identifiable information.

Extracting Training Data from Diffusion Models:

1. The definition of extraction is intentionally conservative as compared to what privacy concerns one might ultimately have. For example, if we prompt Stable Diffusion to generate “A Photograph of Barack Obama,” it produces an entirely recognizable photograph of Barack Obama but not an near-identical reconstruction of any particular training image.

2. The experiments made the assumption such that you have all access in original image and captions, which is not always the case in real-world scenarios.

Future Research Directions

1. Instead of forbidding the model to generate sensitive information, we can give people different levels of privilege to access the data containing different levels of sensitive information from the model.

2. In evaluating the extracted data, we can use more advanced techniques to automatically validate the extracted data, such as using other specific DNN models to classify the data into different categories.

3. In preventing Gen-AI models from being adversarially exploited, there are several short-term goals and long-term goals that need to be addressed. In short-term, we need(1) a comprehensive view of the attack and defense landscape for these technique, (2)comprehensive analysis of the code-related capabilities of LLMs, (3) ensure correctness of LLMs' generated code and (4) a repository of SOTA attacks on various defense techniques. In long-term, we need to (1) bridge the social-technical gap, (2) develop multiple lines of defense for GenAI models, (3) reduce the barrier to entry for GenAI research, and (4) ground LLMs in order to stop hallucinations.

Summary of Class Discussion

Q: Since the data is already public, why should we still be concerned if the training data examples can be extracted from the model? A: The models(GPT2, StableDiffusion) studied in the research paper are trained on public data for the sake of convenience and compliance with the privacy laws. However, production models are often trained on private data, which may contain sensitive information. If the training data can be extracted from the model, it would be a big concern for the privacy of the data.

Q: Companies selling/buying data to LLM, would it be a concern? A: Yes. Google is buying a lot of data from other companies. If the data is leaked for malicious purpose, it would be a big concern.