Homework 4 – Due Tuesday, Dec. 9 at 11:59 PM

Reminder Collaboration is permitted, but you must write the solutions *by yourself without assistance*, and be ready to explain them orally to the course staff if asked. You must also identify your collaborators. Getting solutions from outside sources such as the Web, generative AI, or students not enrolled in the class is strictly forbidden.

1. (**Extracting Memorized Secrets**) *This problem is intentionally open-ended and you may not succeed at the specific task I am asking you to undertake. That's ok! I am more interested in your ideas and your process than in the final result.*

   I finetuned GPT-2 on a small corpus of public domain documents, interspersed with some "secrets" – a combination of snippets consisting of fictitious names, email addresses, keys, and passphrases; plus some longer form "out of distribution" text passages. Your job is extract these memorized secrets out of the model, either partially or completely.

   The finetuned model is contained in the (compressed) folder `gpt2_secret_model` and starter code implementing a baseline extraction strategy is provided in `baseline.py`. You can use any strategy you want to try to locate the hidden secrets, including combining automated steps with manual processing. You may find it useful to take inspiration from the paper `https://arxiv.org/abs/2012.07805`, whose authors made their code available at `https://github.com/ftramer/LM_Memorization` (as always, just please remember to cite your sources!).

   If it's helpful to you, e.g., for performing membership inference, the file `clean_eval.txt` contains text from the same documents used for fine-tuning, without any planted secrets.

   I planted 4 distinct secret snippets with following syntax (with items in square brackets appropriately instantiated):

   ```
    BEGIN CONFIDENTIAL ID: [alphanumeric ID]
   Name:  [First] [Last]
   Email:  [first].[last]+[alphanumeric string]@example.com
   API-Key:  sk_live_[alphabetic string]
   Passphrase:  CANARY::[adjective]-[animal]-[number]
   END CONFIDENTIAL
   ```

   I also planted 4 longer passages with the syntax:

   ```
    BEGIN PASTA ID: [alphanumeric ID]
   [Free Text]
   END PASTA
   ```

   Please submit any code you write in the separate dropbox on Gradescope. In the written part of your assignment:

   - Describe the process you took (i.e., the combination of algorithmic steps and any manual processing you used) to search for the hidden secrets and validate them.
   - What hidden secrets did you find? Did you find any other apparently "memorized" information beyond what I specifically planted?

- Briefly comment on the ethics of this experiment. While both the foundation and finetuned GPT-2 models are trained on "public" data scraped from the Web, does carrying out this experiment raise any privacy concerns? Should we be trying to implement these attacks, and if so, how should we deploy them responsibly?

2. (**Differential Privacy Prevents Reconstruction Attacks**) In this problem, you will show that in a simple data analysis model, it is impossible to reconstruct a sensitive dataset from the output of a differentially private algorithm. In this model, a dataset $b \in \{0,1\}^n$ is a bit vector where bit b_i is person i's information. An algorithm $M : \{0,1\}^n \to Y$ is (ε, δ)-differentially private if for all b, b' differing in a single bit, and all sets of outputs $T \subseteq Y$,

$$\Pr[M(b) \in T] \leq e^{\varepsilon} \Pr[M(b') \in T] + \delta.$$

Say an algorithm $R : Y \to \{0,1\}^n$ is an (α, β)-*reconstructer from* M if for every dataset $b \in \{0,1\}^n$,

$$\Pr[\|R(M(b)) - b\|_0 \leq \alpha n] \geq 1 - \beta.$$

That is, given the output $M(b)$, the algorithm R can reconstruct all but an α *fraction* of the input dataset b with high probability.

(a) Let $i \in [n]$. Define a pair of random variables $(B, B') \in \{0,1\}^n \times \{0,1\}^n$ jointly distributed as follows: Sample B uniformly at random from $\{0,1\}^n$. Set $B'_j = B_j$ for every $j \neq i$ and set B'_i to be an independent uniformly random bit. Show that for every (possibly randomized) algorithm $A : \{0,1\}^n \to \{0,1\}$,

$$\Pr_{\text{coins}(A),(B,B')}[A(B') = B_i] = \frac{1}{2}.$$

(b) Let $A : \{0,1\}^n \to \{0,1\}$ be an (ε, δ)-differentially private algorithm. Let random variable B denote a uniformly random bit vector in $\{0,1\}^n$. Show that for every $i \in [n]$,

$$\Pr_{\text{coins}(A),B}[A(B) = B_i] \leq e^{\varepsilon}/2 + \delta.$$

(c) Suppose $M : \{0,1\}^n \to Y$ is (ε, δ)-differentially private and that $R : Y \to \{0,1\}^n$ is an (α, β)-reconstructor from M. Show that there exists an (ε, δ)-differentially private algorithm A such that for a uniformly random dataset B and a random index i,

$$\Pr_{\text{coins}(A),B,i\in[n]}[A(B) = B_i] \geq 1 - \alpha - \beta.$$

That is, there is a differentially private algorithm that can guess a *random* index of its random input with high probability.

(d) Combine parts (b) and (c) to show that if $M : \{0,1\}^n \to Y$ is $(\varepsilon = 0.1, \delta = 0.1)$-differentially private, then an $(\alpha = 0.1, \beta = 0.1)$-reconstructor from M cannot exist.