CS685 Quiz 3: *attacking AI-generated text detectors*
Released 4/21, due 4/28 on Gradescope (please upload a PDF!)
Please answer each question in 2-4 sentences.

1. Let's say OpenAI releases a plagiarism detection service to verified teachers that allows them to check whether or not a student's submission was generated by ChatGPT. The service compares a candidate submission to every piece of text that ChatGPT has ever generated, and returns a **1** if the submission very closely resembles one of the generations and a **0** otherwise. Now assume one of the teachers is malicious and wants to abuse this service to perform membership inference attacks [*read/skim the paper to understand the idea behind these attacks!*]. First, explain how this teacher might set up their attacks, and what kind of information they might be able to access.

ANSWER

The malicious teacher can set up a membership attack in the following ways

- Prompt the OpenAI API to answer to the questions that he/she has set up for an exam before the students appear for the exam and choose the best possible suitable answer. Now, some students will write the answers to the questions, and as the answers are already in the ChatGPT earlier response, it will flag it a submission that closely resembles one of the generations.
- Also, he/she can ask questions that are generally asked by people and can expect that the answers of his/her student will match one of the previous generations.

2. What are some counter-measures OpenAI can take to prevent the malicious teacher's attacks from succeeding?

- OpenAI can consider the region around the school as a potential data point to see if the teacher has already the asked the questions beforehand. But it may lead to lot of false positives.
- OpenAI can only consider the generations after the exam starts, based on geolocations.

AI Disclosure

AI1: Did you use any AI assistance to complete this homework? If so, please also specify what AI you used.
Your answer here

(only complete the below questions if you answered yes above)

AI2: If you used a large language model to assist you, please paste *all* of the prompts that you used below. Add a separate bullet for each prompt, and specify which problem is associated with which prompt.

- *Your response here*

AI3: (*Free response)* For each problem for which you used assistance, describe your overall experience with the AI. How helpful was it? Did it just directly give you a good answer, or did you have to edit it? Was its output ever obviously wrong or irrelevant? Did you use it to get the answer or check your own answer?

- *Your response here*