Assignment 4 (14 points)

CSE 256: Statistical NLP: Spring 2023 University of California, San Diego
Released: June 2, 2023

Due: June 10, 2023 at **10pm**

In this assignment, we will read this paper and answer a few questions.

Note: Submit your report on Gradescope.

(2 points)

What challenges might we encounter if we try to apply a probability-based detection approach, such as the one proposed by Mitchell et al. in this paper, to the output of models that are not downloadable, and are only accessible via APIs, such as GPT-4? *(A sentence or two should be sufficient here)*

(10 points)

We can always change the LLM so that its output is no longer detectable by a given detection approach. Describe how you would change GPT-2 or any LLM so that its output will no longer be detectable by the Mitchell et al. approach. Your answer should be more specific than "I will change the training" or "I will change the decoding strategy". Be specific about exactly what you will change. Feel free to use illustrative elements in your response, such as examples, equations or pseudo-code. *(A short paragraph is sufficient here, but you are welcome to elaborate.)*

(2 points)

Describe ways in which you (and more generally students) should be able to use large language models (LLMs) in your courses in a productive way. That is, LLMs should not interfere with learning outcomes. Here is what ChatGPT has to say. Provide your own thoughts.
(A couple of sentences should be sufficient here)