CS 839: FOUNDATION MODELS
HOMEWORK 3

Instructions: Read the two problems below. Type up your results. Submit your answers in two weeks (i.e., **Nov. 12th**, end of day). Similarly to last time, you will need a machine for this assignment, but a laptop (even without GPU) should still work.

1. Extracting Training Data From Language Models. One important area in language model security is *attacks meant to extract training data* from fixed, pretrained models. The ability to do this is problematic for several reasons: (1) much of the training data could contain private or personal identifiable information (PII) that is unsafe or even unlawful to reveal, (2) the data could be proprietary. In this problem, we will work on understanding and implementing such attacks.

- 1. *Manual.* Based on our knowledge of general resources that ChatGPT was trained on (wikipedia articles, public domain books, etc.), choose a particular piece of text you hypothesize is in the training set. Report on this phrase. Attempt to force ChatGPT to output it. Report your prompt and whether you were successful.
- 2. *Automated Techniques.* Look at papers such as Ishihara '23 (https://aclanthology.org/2023.trustnlp-1.23.pdf) or Yu et al '23 (https://arxiv.org/pdf/2302.04460.pdf). Select a method that you think would be a good fit for your example from part 1. Describe how you would use it. *Extra credit*: Implement one of these techniques and execute it. Note: this may require using models where you have more access than with ChatGPT, i.e., GPT-Neo and other open-source variants.

2. Multimodal Model Limitations Use a multimodal instruction model like Gemini or GPT 4o. Find a simple visual reasoning problem that stumps the model. You may want to use papers like https://arxiv.org/abs/2403.13315 for inspiration, but craft your own example. Provide the setup, the expected answer, and the model's incorrect answer. Describe your intuition for the reason for the model's failure.