Final Project Update 2

Goals:

1. Have a much more defined scope: (Thursday - Sunday)
    a. Could this be an idea? - Developing a model resistant to common AI detecting evasion techniques, such as paraphrasing, adding noise, or using prompt engineering, using adversarial training. (could be different idea, maybe just creating our own model to detect AI generated text)
    b. Try to find similar work to get more of a background

2. "Data (required) (Thursday - Sunday) -- A thorough description of the data you will use (eg. the source of the data, your evaluation of the quality of the data while doing EDA and any changes, transformations or features you engineered on the data)"
    a. Find 1-2 datasets each, perform EDA on them.
        i. Dataset #1 - AI Generated Essay and Human Written Essay
        ii. Dataset #2 - AI detection dataset (1.39M rows)
        iii. Dataset #3 - largest and most comprehensive dataset for AI-generated text detectors

3. Create another Google slides presenting all our scope, more background info, datasets, EDA, etc… (Sunday-Monday)

4. Record presentation on Zoom (Monday)

5. Submit to Canvas (Monday)
    a. What do we have to submit?
        i. Recording of the presentation, Group Contract

Links - taiyo:

https://arxiv.org/search/cs?query=AI-generated+content+detection&searchtype=all&abstracts=show&order=-announced_date_first&size=50

https://arxiv.org/abs/2301.11305

https://gptzero.me/

https://github.com/openai/gpt-2-output-dataset

https://www.kaggle.com/competitions/llm-detect-ai-generated-text

https://huggingface.co/datasets/Hello-SimpleAI/HC3