**Your presentation has contributed to this grade with the following breakdown:**

**\* Originality/Existing Market Usage/External Validation Concept**

**\* Real World Application of the Project Concept**

**\* Overall Presentation: Introduction/Data explanation/Architecture and model comparisons/Results quality**

**\* Model Justifications and result visualization**

**Submit a final report of your project with the following:**

**\* Introduction/Literature review/related work**

**\* Data exploration and processing/Problem Formulation and model selection**

**\* Evaluation, result analysis, visualization**

Notes/Readings

Lu, N. (2023, May 18). *Large Language Models can be Guided to Evade AI-Generated Text Detection*. arXiv.org. <https://arxiv.org/abs/2305.10847>

It reveals limitations of current AI text detectors through SICO, while also providing a reliable methodology to evaluate detector robustness. The results motivate development of more advanced detectors that are robust to these evasion attacks.

Few points from this paper "Large Language Models can be Guided to Evade AI-Generated Text Detection" :

* It proposes a novel method called Substitution-based In-Context example Optimization (SICO) to automatically generate prompts that can guide LLMs like ChatGPT to generate human-like text that can evade detection by current AI text classifiers.
* One main advantage of SICO over paraphrase attacks is that SICO can make LLMs directly generate text which cannot be detected, thus eliminating the need for an extra paraphrasing step. This can completely avoid possible semantic drift or textual quality degradation caused by paraphrasing.
* It comprehensively evaluates SICO against 6 current state-of-the-art AI text detectors over 3 real-world text generation tasks - academic essay writing, open-ended QA, and fake review generation.
* The results demonstrate that SICO enables ChatGPT to successfully evade all 6 detectors, causing their AUC to drop by an average 0.54, essentially making them perform worse than random classifiers. This reveals the vulnerability of current detectors.
* The approach is efficient, requiring around 500K tokens to train prompts that can evade detectors. It is also flexible and transferable to other LLMs.