Context:

“At the forefront of academic concerns about LLMs is their potential to enable plagiarism.”[1] With the rise in LLMs, their availability, and their ability to write high level responses, it becomes important to have a reliable tool to detect AI generated text.

Criteria for Success:

This tool would be useful for anyone reviewing submitted work, like teachers and professors, as well as the general public. With the excess of new media uploaded everyday, it’s important to know if the information is reliable and genuine. For this reason, I believe a minimum of 90% accuracy would be needed to trust an AI detect model. In the instance that a professor uses this tool on a student’s submitted essay, that student could receive a 0 or a mark on their record for cheating if the essay is deemed AI written. The consequence if incorrect would be severely unfair.

Scope of Solution Space:

Final product will be a classifier model that intakes text, in the form of an essay, and outputs a percent chance whether the text was written by AI or not.

Constraints:

Time. As this will be entered into a competition, a submission must be ready by the deadline. While there might be options available during training that would greatly improve accuracy, time used during training must be taken into account.

Stakeholders:

- The Learning Agency Lab, as the Kaggle competition host.
- Eleanor Thomas, Springboard Mentor, for capstone grading.

Data Sources:

“Augmented Data for LLM” by Jonathan Herrera
https://www.kaggle.com/datasets/jdragonxherrera/augmented-data-for-llm-detect-ai-generated-text