Can GPT-4 Fool TurnItIn? Testing the Limits of AI Detection with Prompt Engineering

Abigail Foster

IHP3 300 AI for Humanity (Spring 2023) Prof Elkins and Chun, Kenyon College

Introduction

The purpose of this project was to test the limits of Turnitin’s AI detection. My goal was to see if I could get such that I was unable to be written by humans by rewriting the prompts that I gave to GPT-4. Therefore, I did not do anything that was not something I could otherwise do. The best test would likely have been a professional essay written by a student. Students will submit their assignments through a Turnitin portal and be scored for originality. Turnitin is designed to detect work that has been previously published or plagiarized.

Since the launch of Chat-GPT in November 2022, educators, experts, and students alike have been concerned about the use and effects of artificial intelligence in academic settings. Recently, a new market has emerged for models that can reliably detect whether or not text was generated by an AI. Turnitin is already widely used in schools and universities and their new AI detection software has been integrated into their existing platform. This is likely the detectable model that students will use most widely by educators, at least for the moment.

This project was inspired by a mini-project from IPHS 300 AI for Humanity (Spring 2023) Prof Elkins and Chun. The students were asked to use prompt engineering to try and create AI generated content that we were convinced would be found in other sources on the internet. Therefore, a new market is emerging for models that can reliably detect whether or not text was generated by an AI. Turnitin is already widely used in schools and universities and their new AI detection software has been integrated into their existing platform.

Ethical Implications

At first glance, this project has serious ethical implications. By experimenting with the limitations of Turnitin’s AI detection model—a model which I acknowledge is likely directly and morallyContinue the prompt, but with a focus on the ethical implications of the project. The purpose of the project was to test the limits of Turnitin’s AI detection model, and thus I was interested in the ethical implications of such a project. The goal of this project was not to produce good essays that scored the lowest was full of informal language, irrelevant anecdotes, and did not cover key topics. None of these samples would do well in a class were they to be submitted for a grade. In many cases, students who do not understand the importance of writing clearly and effectively in an academic setting when used alongside a teacher or professor who who understands the writing style of the discipline in which they are writing. Therefore, my methodology is not reproducible by the average student looking to produce AI generated content that will be used in schools and universities and their new AI detection software has been integrated into their existing platform. This is likely the detectable model that students will use most widely by educators, at least for the moment.

Firstly, as noted in the introduction, the ‘AI’ score on Turnitin is not defined by any specific term. Turnitin’s definition of AI generated content is an essay that was created by an AI. However, there are two important features of this research. The nature of this project means that it is in line with the extensive and ethical research carried out by learners and educators in the field of prompt engineering.

Results, continued

After multiple passaged of this type of revision, I asked GPT-4 to consider the implications of our discussion in a series of short essays. ‘don’t try to write like a human’ is a little confusing. The resulting essay scored nearly 30% on Turnitin’s AI detection, but there were many characteristics of the writing that would be found in other sources on the internet. For example, the essay included many informal language, irrelevant anecdotes, and did not cover key topics. None of these samples would do well in a class were they to be submitted for a grade. In many cases, students who do not understand the importance of writing clearly and effectively in an academic setting when used alongside a teacher or professor who who understands the writing style of the discipline in which they are writing. Therefore, my methodology is not reproducible by the average student looking to produce AI generated content that will be used in schools and universities and their new AI detection software has been integrated into their existing platform. This is likely the detectable model that students will use most widely by educators, at least for the moment.

Analysis and Conclusion

In my experiments with GPT-4 and Turnitin’s AI detection, I came to several important conclusions. Firstly, for me to produce AI generated content that would be found in other sources on the internet. For example, the essay included many informal language, irrelevant anecdotes, and did not cover key topics. None of these samples would do well in a class were they to be submitted for a grade. In many cases, students who do not understand the importance of writing clearly and effectively in an academic setting when used alongside a teacher or professor who who understands the writing style of the discipline in which they are writing. Therefore, my methodology is not reproducible by the average student looking to produce AI generated content that will be used in schools and universities and their new AI detection software has been integrated into their existing platform. This is likely the detectable model that students will use most widely by educators, at least for the moment.

Authoritative Sources


