

The Limits of AI Content Detectors

Hongyu Wu¹ and Tom Flanagan^{1#}

¹Singapore American School, Singapore

#Advisor

ABSTRACT

As ChatGPT became a popular and powerful language model used by people worldwide in 2023, the problem of students using it to cheat on schoolwork became palpable. While many existing AI content detectors can detect AI-generated texts, such as GPT-2 Content Detector and GPTZero, the accuracy of an AI content detector in detecting generated essays that have been post-edited by humans is unknown. This research discovered the limitations of the GPT-2 Content Detector and answered the question, “How does human post-editing of AI-generated high school English essays affect the result of an AI content detector?” Ten English essays were generated using ChatGPT Plus based on prompts from high school English teachers. Each essay was then edited in 5 different ways to create pairs of unedited and edited essays. All unedited and edited essays were evaluated using GPT-2 Output Detector Demo, and then the results from the detector were studied and analyzed. It was found that introducing spelling mistakes in generated essays and processing the essays with QuillBot will make the result of AI content detectors less accurate. The findings from this research can be used as a guide for companies developing AI-generated text detectors, making them more accurate when dealing with edited generated text. The findings can also be helpful for schools and educators, because knowing that students can edit essays to bypass AI content detectors, educators can develop new ways to examine students’ writing ability.

Introduction

The recent breakthrough in Artificial Intelligence (AI) has led to the development of many Large Language Models (LLMs), one example being ChatGPT developed by OpenAI. LLMs can be simply defined as computer programs that can understand and generate human-like language, and ChatGPT is an LLM that can remember, understand (Bommarito et al., 2023), and engage in conversation with users (OpenAI, 2022). However, the availability of these LLMs has also led to concerns about student plagiarism, as students may use them to generate answers to homework assignments and other academic tasks (Winter, 2023).

One potential solution to this problem is the use of AI content detectors, which can analyze text and determine whether it is likely to have been generated by an AI system (Gao et al., 2022). However, the effectiveness of these detectors when it comes to detecting essays that have been post-edited by students is unknown. For example, a student may use an AI text generator to generate a response to an essay question, and then edit the generated text by replacing some of the vocabulary with synonyms and changing the sentence structure. Whether AI content detectors can still accurately detect these edited essays poses an unanswered question.

This study aimed to investigate this issue by examining the impact of post-editing on the accuracy of AI content detectors in detecting AI-generated text. Specifically, this study used the AI text generator ChatGPT Plus developed by OpenAI, this was chosen because while there are many text generators out there, ChatGPT is one of the most popular as it is free for the public to use. This study uses a paid version of ChatGPT called ChatGPT Plus, which allows access to ChatGPT during peak times and generates results faster; however, there are no differences between ChatGPT and ChatGPT Plus other than the calculation speed, stability, and earlier updates. The AI content detector used in this research is GPT-2 Output Detector Demo (GPT-2 ODD), an OpenAI model (Salminen et al.,

2021). The findings of this research will have important implications for the development and implementation of anti-cheating software and for the use of AI-generated text in academic settings.

The research question for this study is: How does human post-editing of AI-generated high school English essays affect the result of an AI content detector?

Literature Review

Generative AI

As generative AI technology emerged, numerous studies have focused on exploring the myriad of AI generation tools available, ranging from Text-to-Image models such as Stable Diffusion to Text-to-Video models like Phenaki (Gozalo-Brizuela & Garrido-Merchan, 2023). However, this research specifically concentrated on Text-to-Text models. The Text-to-Text model employed in this study was ChatGPT, a state-of-the-art language model that had gained over a million subscribers within a week of its launch due to its recent development and popularity (Baidoo-Anu & Owusu Ansah, 2023).

Generative Pre-trained Transformer 3 (GPT-3)

To comprehend ChatGPT, it is essential to examine GPT-3, released in 2020, as ChatGPT utilizes the same architecture. GPT-3, developed by the Silicon Valley research firm OpenAI, is a potent natural language processing (NLP) system. It is a third-generation language model that employs deep learning to generate human-like text. Essentially, it is a computational system designed to generate sequences of words, code, or other data based on human input, known as the prompt (Floridi & Chiriatti, 2020). GPT-3 can respond to virtually any topic and generate related original text content that is challenging to distinguish from human writing (Dehouche, 2021). In fact, one study suggests that GPT-3 writes better than many people (Elkins & Chun, 2020).

In addition to GPT-3's ability to produce human-like text, a study found that GPT-3 could generate clear and concise descriptions of its own capabilities and features (Thunström et al., 2022). This research also indicated that ChatGPT might possess similar abilities, given that it is based on the GPT-3 architecture. One specific research paper identified that one of GPT-3's models, Text-davinci-003, approached human-level performance on remembering, understanding, and applications (Bommarito et al., 2023). However, although GPT-3 is powerful in generating human-like texts, there are areas where it does not perform as well. One particular version of the GPT-3 model, Text-davinci-003, underperformed on numeric reasoning in zero-shot prompts (prompts it has never been trained on before) (Bommarito et al., 2023).

Given GPT-3's performance differences when it came to completing different types of tasks, debates arose regarding the ability of GPT-3. While some individuals believed that GPT-3 was a powerful NLP (Dehouche, 2021), others argued that it was not the beginning of general artificial intelligence (Floridi & Chiriatti, 2020) and had shortcomings in mathematical, semantic, and ethical problems (Elkins & Chun, 2020).

ChatGPT

Following GPT-3, ChatGPT completed its training in early 2022, it is a large language model fine-tuned from a model in the GPT-3.5 series (OpenAI, 2022). Much like its predecessor, ChatGPT is capable of understanding complex concepts, adjusting its speech based on the audience, and logically analyzing situations (Benzon, 2023). In addition to its powerful abilities, ChatGPT quickly gained popularity as its language models were made publicly available (Aydn et al., 2023).

ChatGPT in Education

While ChatGPT is powerful, one of the concerns it raised is that it can be used to cheat on assessments (Winter, 2023). Given ChatGPT's availability to the public and its capability for logical reasoning, students can easily exploit this tool, posing a potential threat to the integrity of online exams (Susnjak, 2022). The existence of this academic integrity threat was subsequently confirmed in a study conducted in the Netherlands, where researchers had ChatGPT take an English reading test intended for high school students. Notably, ChatGPT received a grade of 7.18, equivalent to the average grade of students in the nation, indicating its exceptional performance on the test (Winter, 2023). Similar studies have also highlighted the potential risks ChatGPT poses to the integrity of essay submissions, particularly in higher education settings (Ventayen, 2023).

Ventayen's research revealed that the AI model fails to accurately acknowledge sources or citations (Ventayen, 2023), meaning that if students use ChatGPT, the generated work could be susceptible to plagiarism. Furthermore, the study noted that students can utilize QuillBot, a paraphrasing tool, to rephrase AI-generated content, creating an illusion of plagiarism-free work. Subsequently, students can employ the AI-generated content and falsely assert ownership of the ideas presented (Ventayen, 2023).

Detection of AI-generated Texts

Given the identified issues with academic integrity caused by AI, it is crucial to conduct research aimed at addressing this situation. One study emphasized the significance of continually developing well-thought-out and meticulously researched solutions (Popenici & Kerr, 2017).

One solution suggested by a study is the use of a content detector when identifying AI-generated text (Salminen et al., 2021). The content detector would be capable of identifying whether a piece of text is produced by AI, meaning educators could potentially use the detectors to identify academic integrity breaches.

Currently, various methods exist for identifying AI-generated text. In one study that aimed to distinguish machine-written text from human-written text, a feature-based classifier was employed. This classifier considered features such as text length, word frequency, and the presence of specific patterns or structures (Fröhling & Zubiaga, 2021). Another study employed a different detection method, utilizing machine learning to construct a binary classifier (Lavoie & Krishnamoorthy, 2010).

In addition to the various identification methods, there is an existing RoBERTa-based sequence classifier called GPT-2 ODD, an AI output detector that is an OpenAI model (Salminen et al., 2021). This detector provides an abstract score of the "fakeness" of a piece of text, ranging from 0.02% to 99.98% "fake," with higher scores indicating that the text was more likely produced by an AI algorithm.

Method

To explore the effect of human post-editing AI-generated high school English essays on the result of an AI content detector, the researcher used a quantitative method to evaluate the result of the AI content detector. This is because this experiment collects purely quantitative data and does not involve surveys. An experimental research method was used because it allows for a controlled environment in which the independent variable (different ways to post-edit the essays) can be manipulated and the dependent variable (results of the AI content detector) can be measured and analyzed. The utilization of a controlled experimental design aligned with the study's research objective, as it is possible to isolate and identify cause-and-effect relationships between the independent and dependent variables. In addition, to address concerns about academic integrity (Ventayen, 2023) in this research, ChatGPT Plus was programmed to produce text in the form of school essays.

Preparation

Teachers and Essay Prompts

This research used probability sampling to collect the essay prompts where each high school English teacher had an equal chance of being selected. To prepare for the research, the researcher used a random sampling method to select ten random teachers out of all high school English teachers at Singapore American School for different high school essay prompts. Out of all the essay prompts received, five were then randomly selected again to be used in the research. Some of the essay prompts received from teachers are from the official AP English Language and Composition (AP Lang) exams, each AP Lang exam contains three questions. Since the first question in the AP Lang exam usually contains graphs and the current version of ChatGPT does not have the ability to process graphs, questions two or three were randomly selected again for prompt generation. Probability sampling was suitable for this research because it minimized the risk of selection bias. It also maximized the chance that the essay prompts used in this research were commonly used by high school students.

Software

Three online software were used in this research, two of which are free for public use. The first software is ChatGPT Plus, developed by OpenAI, the popular AI language model capable of generating compelling and accurate answers (Susnjak, 2022). ChatGPT is free and available for public use; however, this research uses ChatGPT Plus, a paid version, that allows access to ChatGPT during peak times and generates results faster (ChatGPT, 2023). There were no differences in terms of results generated by ChatGPT and ChatGPT Plus. The second software is an OpenAI model called GPT-2 ODD (Salminen et al., 2021), which is an AI content detection tool. It is a RoBERTa-based sequence classifier that rates texts as "fake" or "real" with scores ranging from 0.02% to 99.98%. A higher score indicates that the text was more likely produced by an AI algorithm (Gao et al., 2022). GPT-2 ODD was chosen over other detectors such as GPTZero because it was a more mature model and was used across multiple existing research papers, and also because it provides an exact estimation of the likelihood that a piece of text was generated. Compared to newer detectors like GPTZero, which only provide a generalized description of the likelihood that the text was generated, GPT-2 ODD is more suitable for statistical analysis. Moving on, the third software is an online paraphrasing tool called QuillBot made by Rohan Gupta. It is capable of rewriting sentences with synonyms and different sentence structures. This study used QuillBot as a method to manipulate AI-generated texts.

Procedure

Essay Generation

With a similar approach from a study about the GPT-2 Output Detector (Gao et al., 2022), the researcher evaluated the essays generated by ChatGPT Plus with 50 different post-edited essays. After randomly selecting five essay prompts provided by the high school English teachers, the researcher inputted these five prompts into ChatGPT Plus to generate two essays under each essay prompt. Two essays were generated under each prompt to minimize the impact of the different results in ChatGPT Plus generation; ChatGPT Plus provides different results each time it generates, and some results might be more easily detectable; generating two essays under the same prompt can minimize this variation.

To ensure consistency in the type and length of essays generated by ChatGPT Plus, and drawing inspiration from an existing study, a uniform prompt template was utilized (Susnjak, 2022). The template is presented below:

Please answer the following prompt in several paragraphs using 500 words with examples and supporting arguments: [English prompt here]

Some of the essay prompts require additional information, such as reference passages or the name of a novel (see Appendix A to E). For these types of essay prompts, the following uniform prompt template was used:

[Additional information here]

According to the above information, please answer the following prompt in several paragraphs using 500 words with examples and supporting arguments: [English prompt here]

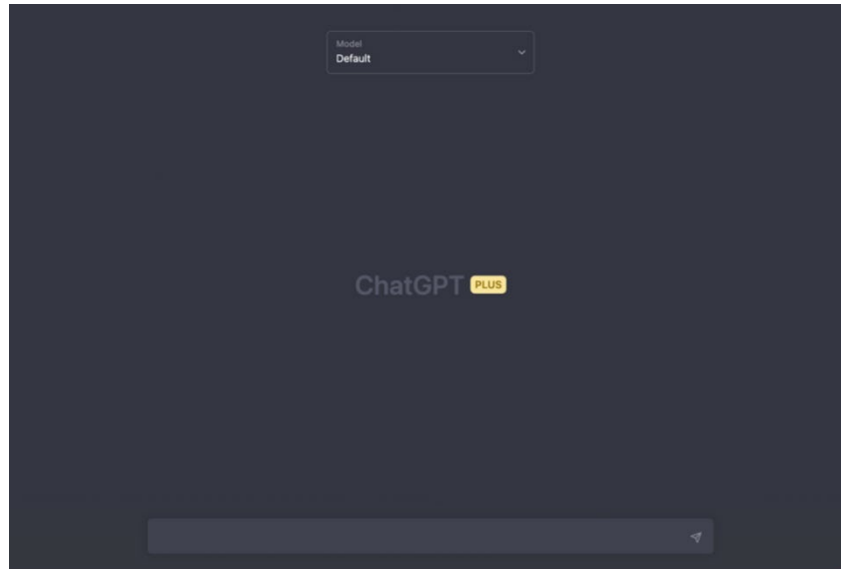


Figure 1. ChatGPT Plus’s Website Interface with the Text Input Prompt at the Bottom (ChatGPT, 2023).

Editing of Essays

Each of the generated essays was edited using each of the following five techniques:

Use of synonyms: One noun was selected from each paragraph and replaced with its synonym. This was used as one of the essay editing techniques because it is one of the more straightforward edits that students can conduct while trying to bypass an AI content detector.

- After entering the noun, the researcher replaced the third noun in each paragraph of each generated essay with the first synonym shown on the website Thesaurus.com at <https://www.thesaurus.com>.

Introduction of spelling mistakes: One spelling mistake was introduced in each paragraph. This was used as an editing technique because ChatGPT Plus is built to generate essays with good grammar and spelling, so there is a possibility that an AI content detector would not be able to detect an essay that is AI generated if it has spelling errors.

- In one of Dr. Graham Rawlinson’s researches, he discovered that the order of the letters in a word is not as important when it comes to recognizing a word. However, the correct position of the first and last letters plays a more significant role (Rawlinson, 1976). When the researcher introduced misspelled words in this experiment, the first and last letters of the words remained unchanged, and the positions of two random middle letters were swapped.
- Specifically, a random word with more than three letters in each paragraph of each generated essay was modified with a spelling error. The noun must be more than three letters because a three-letter word only has one middle letter and therefore cannot be swapped with anything else.

Introduction of grammar errors: In research conducted by Dr. Gino G. Sumalinog, he identified that one of the common grammatical errors that high school students make is using the incorrect subject and verb agreement, specifically the misuse of “has” and “have” (Sumalinog et al., 2018). In this research, grammatical errors were introduced by replacing the first “has” with “have,” and the first “have” with “has” in each paragraph. In the case where a paragraph does not contain “has” or “have,” the first “is” was replaced with “are,” and vice versa. If a paragraph does not contain “is” or “are” as well, no modification was made to the paragraph.

Introduction of formatting errors: An extra space was added in a random location (between two words) in the essay. The location was generated by a random number generator, with the maximum number being the word count of the essay. The formatting error was introduced in essays because ChatGPT Plus was built to generate essays with good spelling and grammar; therefore, it is unlikely that ChatGPT Plus will generate essays with an extra space between words. If an AI content detector is not built to compromise errors in writing, then adding a formatting error might drastically affect the result of the content detector.

Paraphrasing the entire essay using QuillBot: The researcher inputted the essay into Quillbot and asked it to paraphrase it under the “Fluency” mode with the maximum synonyms available for free users. This study used the “Fluency” mode because it ensures that the edited text is readable and error free (QuillBot, 2022). QuillBot was used as one of the modifications in this research because it replaces words with synonyms and reorders the sentences. QuillBot is also a free-to-use tool for the public, so it is an easily accessible modification that students can use in their writing.

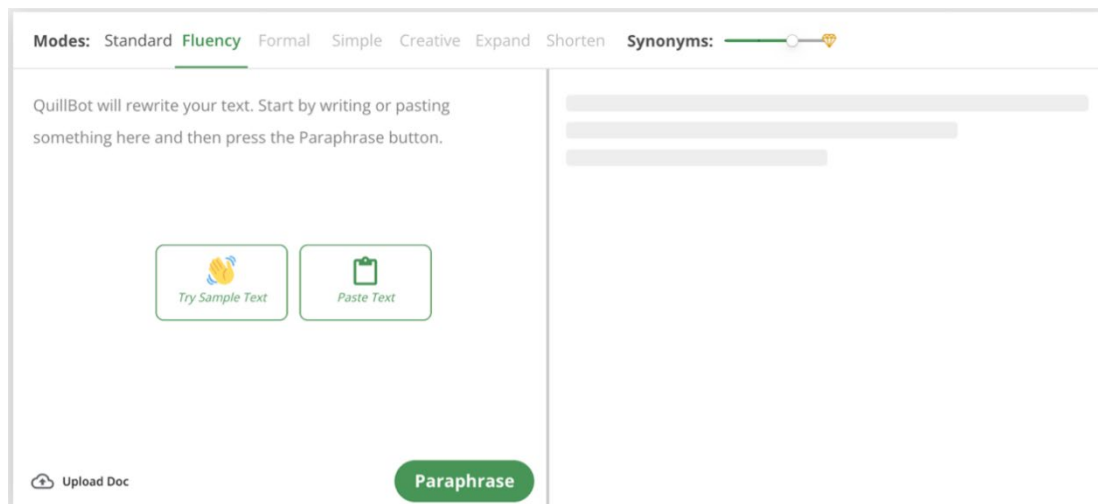


Figure 2. QuillBot’s Publicly Available Website Interface with the Text Input Prompt on the Left.

Evaluating Essays

All unedited and edited generated essays were evaluated using GPT-2 ODD, and the results were recorded.

GPT-2 Output Detector Demo

This is an online demo of the GPT-2 output detector model, based on the [huggingface/transformers](#) implementation of RoBERTa. Enter some text in the text box; the predicted probabilities will be displayed below. The results start to get reliable after around 50 tokens.

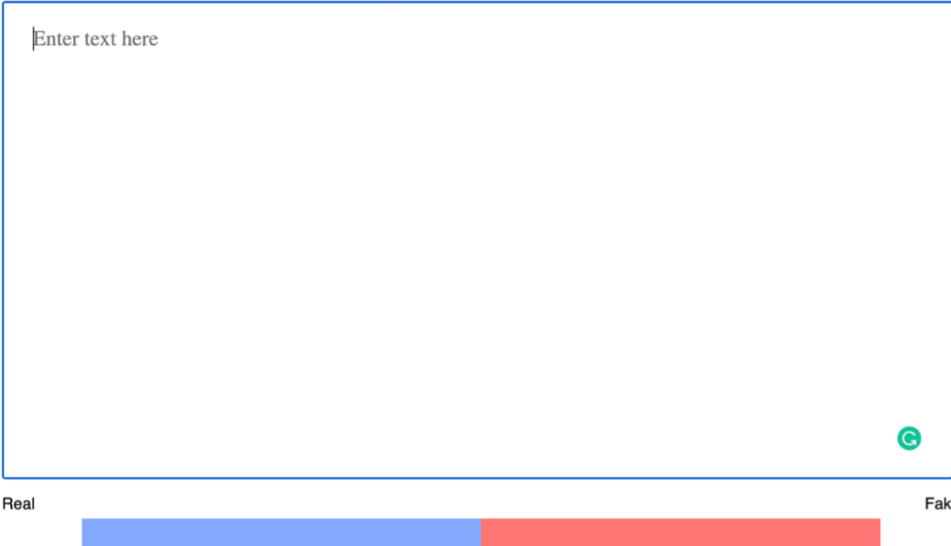


Figure 3. GPT-2 Output Detector Demo's Publicly Available Website Interface.

Data Analysis

The recorded results from the GPT-2 ODD for essays before and after edits were analyzed and graphed to determine the effect of the various editing techniques on the ability of the GPT-2 ODD to detect AI-generated text.

Specifically, for each of the five types of edited essays and the unedited essays, a T-test was conducted to determine whether editing the essay had an impact on the accuracy of the output detector. A p-value was calculated to assess the significance of the results. A T-test was chosen as it is suitable for analyzing whether there is a significant effect of a categorical Independent Variable (different ways to post-edit the essay) on a numeric Dependent Variable (accuracy from the AI output detector). Moreover, taking a similar approach to existing research on the accuracy of the GPT-2 output detector (Gao et al., 2022), the median scores obtained from the AI output detector were compared for the edited and unedited essays to determine if the AI output detector was more accurate for one or the other, and to what extent.

Results

The researcher used experimental research to explore the impact of human post-editing of AI-generated high school English essays on the result of an AI content detector. The independent variable for this research was the different ways to post-edit the essays, and the dependent variable was the results of the AI content detector (GPT-2 ODD). Ten high school English essays were generated with ChatGPT Plus, and each of these ten essays was human post-edited in five different ways. All the essays were then fed to GPT-2 ODD and the "percentage fake" detected by GPT-2 ODD for each essay (both edited and unedited) were recorded and displayed in Table 1 below. The researcher later used a quantitative method (t-test) to further evaluate the result of the AI content detector.

Table 1 below displayed the descriptive statistics of the "percentage fakes" detected by GPT-2 ODD for each of the unedited essays and its paired edited essays.

Table 1. Table of Descriptive Stats of the Percentage Fake Detected by GPT-2 Output Detector Demo for Each Generated Essay.

	Percentage fake detected by GPT-2 Output Detector Demo					
Essays	Unedited	Synonyms	Spelling Mistakes	Grammar Errors	Formatting Errors	QuillBot
P1 V1 ¹	99.98%	99.70%	74.57%	99.98%	99.96%	99.97%
P1 V2	99.98%	99.98%	40.13%	99.98%	99.88%	22.97%
P2 V1	0.11%	0.06%	0.02%	0.37%	0.02%	0.02%
P2 V2	99.94%	99.95%	0.05%	99.92%	62.72%	88.14%
P3 V1	99.98%	99.97%	0.59%	99.98%	52.47%	0.03%
P3 V2	99.97%	99.98%	4.05%	99.97%	77.14%	0.17%
P4 V1	99.98%	99.98%	0.19%	99.98%	99.98%	99.84%
P4 V2	99.98%	99.98%	3.61%	99.98%	99.98%	99.62%
P5 V1	99.93%	35.91%	0.70%	99.86%	95.47%	0.06%
P5 V2	70.35%	0.45%	0.02%	0.25%	70.35%	0.13%

“P” in the Essays column stands for “Prompt” and “V” stands for “Version.” “P1 V1” means prompt one version one.

From Table 1 there were two outliers for the unedited essays, P2 V1 and P5 V2 both had a lower “percentage fake” than others detected by GPT-2 ODD. The unedited version of P2 V1 only had a “percentage fake” of 0.11%, meaning that GPT-2 ODD could not recognize that this essay is AI generated.

Looking at the data from the edited essays, the “percentage fake” value for essays with spelling errors appeared to be significantly lower than the essays without edits. Furthermore, essays with spelling errors also had the lowest “percentage fakes” when compared to other types of post-editing techniques. For the essays with formatting errors and essays that have been processed with QuillBot, some of the data had lower “percentage fakes” when compared to their unedited versions, while others did not have much difference. It was also interesting to see that post-editing essays by replacing words with synonyms, or introducing grammar errors into the essays did not appear to have a significant impact on the majority of the essays.

In order to answer the research question of how human post-editing of essays impacts the result of an AI content detector, it was important to understand whether post-editing the generated essays had a significant impact on the result of the detector when compared to unedited essays in a larger context. It was necessary to compare and understand the difference in “percentage fakes” detected by GPT-2 ODD between the unedited and edited essays. To do so, the researcher aimed to use data collected to reject the null hypothesis (H_0) in which there is no difference in result measured by an AI content detector for edited and unedited essays, with the hopes of accepting the alternative hypothesis (H_a) in which there is a difference.

Three separate independent-sample t-tests were conducted to compare “percentage fake” measured by GPT-2 ODD in the unedited essays and the essays with spelling mistakes, formatting errors, and essays processed with QuillBot. The significance of implementing a t-test is that a mean difference can be calculated between the groups in hopes of rejecting H_0 .

Table 2. Summary Table for Percentage Fake (%) Detected by GPT-2 Output Detector Demo for Each Generated Essay.

	Sample Size (<i>n</i>)	Mean (M)	Standard Deviation (SD)
Unedited	10	87.02	31.92
Synonyms	10	73.60	43.50
Spelling Mistakes	10	12.39	25.09
Grammar Errors	10	80.03	42.01
Formatting Errors	10	75.80	31.99
QuillBot	10	41.10	48.64
Observations	60		

Table 2 compares the “percentage fakes” across all essays (both unedited and edited) by comparing sample size (*n*), mean (M), and standard deviation (SD). It can be observed that though sample sizes are the same (*n*=10), the mean “percentage fake” measured per sample was significantly lower for essays with spelling mistakes and essays processed with Quillbot than unedited essays ($12.39 < 87.02$, $41.10 < 87.02$). The mean “percentage fake” for essays with synonyms and formatting errors were similar, both with lower mean than unedited essays as well ($73.60 < 87.02$, $75.80 < 87.02$). The mean “percentage fake” for essays with grammar errors had the highest correct detection percentage when compared to other edited essays (80.03). The SD is at lowest for essays with spelling mistakes (25.09), and highest for essays processed with QuillBot (48.64).

By using “percentage fakes” to measure GPT-2 ODD’s accuracy, we can introduce Figure 4 to visualize the mean and median differences in results detected by GPT-2 ODD and understand more deeply about the difference in “percentage fakes” detected for essays with spelling errors and unedited essays.

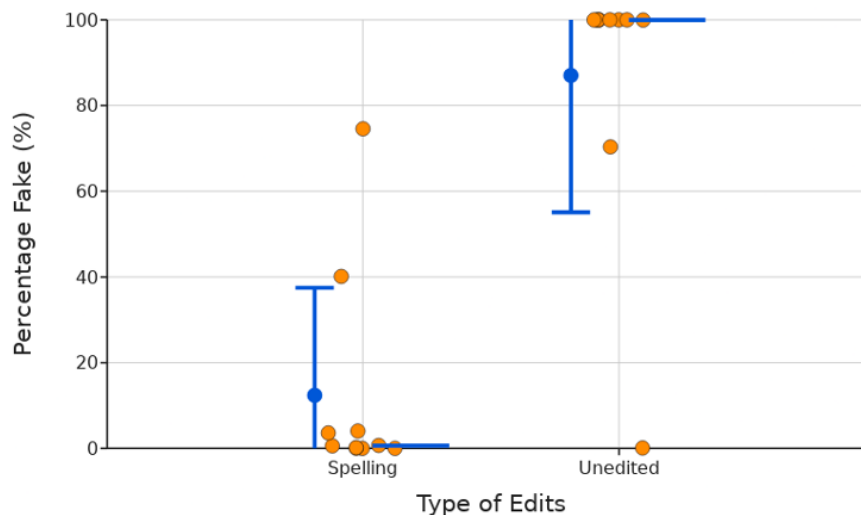


Figure 4. Illustration of the Percentage Fake Detected by GPT-2 Output Detector Demo Between Essays with Spelling Errors and Unedited Essays. Blue dots show the mean. Error bars above and below show the standard deviation. Blue line on the right shows the median value.

Figure 4 demonstrated a difference in “percentage fake” detected for essays with spelling errors and unedited essays. It was visible from Figure 4 that both the mean and median for the essays with spelling errors were significantly lower than those for the unedited essays. There was a significant difference in the scores for essays with spelling errors ($M = 12.39$, $SD = 25.09$) and unedited essays ($M = 87.02$, $SD = 31.92$) conditions; $t(18)=5.8$, $p = 0.01$.

For the research being conducted, the significance level was $p \leq 0.05$. The significance of the p-value (0.01) indicated strong confidence in the H_a , and rejection of the H_o between essays with spelling errors and unedited essays.

Together with the graph and the t-test, they helped to answer the research question by suggesting that editing the essays does in fact had an impact on the result of an AI content detector. In this specific scenario, the edited essay is less likely to be detected by an AI content detector when compared to unedited essays.

Figure 5 was used to visualize the mean and median differences in results detected by GPT-2 ODD for essays with formatting errors and unedited essays.

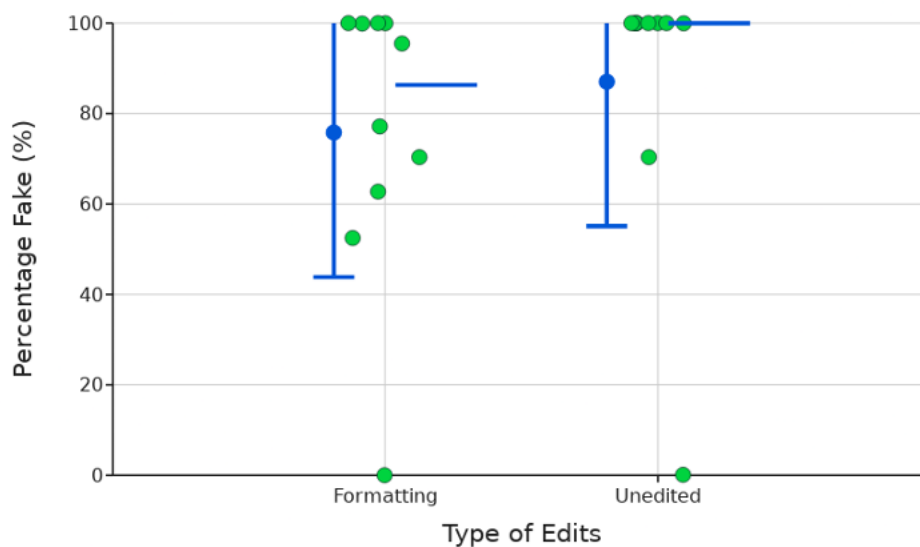


Figure 5. Illustration of the Percentage Fake Detected by GPT-2 Output Detector Demo Between Essays with Formatting Errors and Unedited Essays. Blue dots show the mean. Error bars above and below show the standard deviation. Blue line on the right shows the median value.

Figure 5 demonstrated a difference in “percentage fake” detected for essays with formatting errors and unedited essays. There was not a significant difference in the scores for essays with formatting errors ($M = 75.80$, $SD = 31.99$) and unedited essays ($M = 87.02$, $SD = 31.92$) conditions; $t(18)=0.79$, $p = 0.44$.

The significance of the p-value (0.44) indicated that the researcher failed to reject the H_o between essays with formatting errors and unedited essays.

Figure 6 was used to visualize the mean and median differences in results detected by GPT-2 ODD for essays processed with QuillBot and unedited essays.

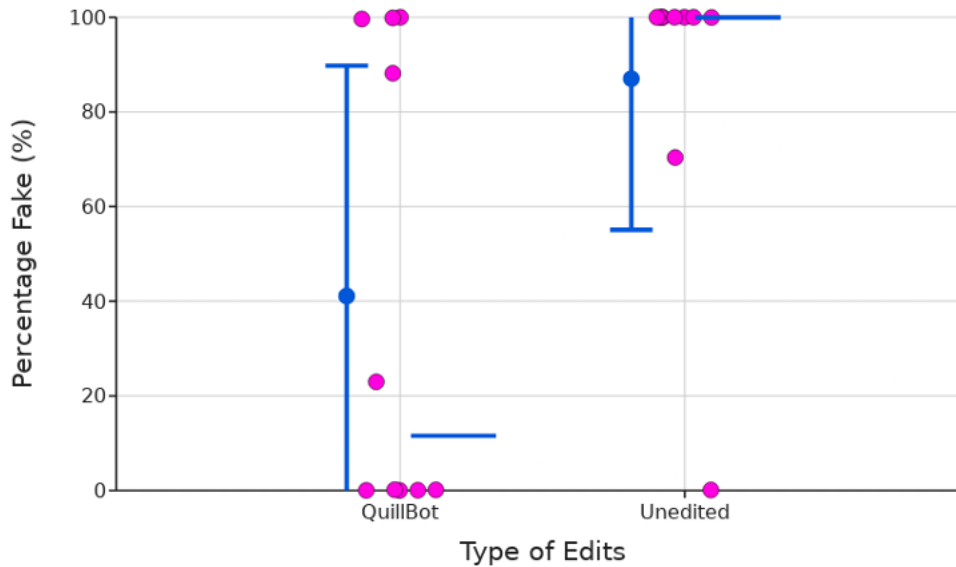


Figure 6. Illustration of the Percentage Fake Detected by Gpt-2 Output Detector Demo Between Essays Processed with QuillBot and Unedited Essays. Blue dots show the mean. Error bars above and below show the standard deviation. Blue line on the right shows the median value.

Figure 6 demonstrated a difference in “percentage fake” detected for essays processed with QuillBot and unedited essays. There was a significant difference in the scores for essays processed with QuillBot ($M = 41.1$, $SD = 48.64$) and unedited essays ($M = 87.02$, $SD = 31.92$) conditions; $t(18)=2.5$, $p = 0.02$. The significance of the p-value (0.02) indicated strong confidence in the H_a and rejection of the H_o between essays processed with QuillBot and unedited essays. After running the three t-tests, it was identified that there were significant differences between the essays with spelling errors, the essays processed with QuillBot with the unedited essays. To identify which type of these two edits had a more significant impact on the result of GPT-2 ODD, the researcher compared the median of the “percentage fakes” of the essays. The median was compared because this research took a similar approach to existing research on the accuracy of the GPT-2 output detector (Gao et al., 2022), since the data collected contains potential outliers, comparing medians can return more accurate answers than comparing means.

Table 3. Comparing Median of the Percentage Fake Detected by GPT-2 Output Detector Demo for Each Generated Essay.

	Sample Size (n)	Median (Mdn)
Unedited	10	99.98
Synonyms	10	99.96
Spelling Mistakes	10	0.65
Grammar Errors	10	99.98
Formatting Errors	10	86.31
QuillBot	10	11.57
Observations	60	

Table 3 indicated that the median of the essays with spelling mistakes and essays processed with QuillBot were significantly lower than the unedited essays ($0.65 < 99.98$, $11.57 < 99.98$), this showed that post-edited essays with the above two techniques are less likely to be detected by GPT-2 ODD when compared to unedited essays. Furthermore, when comparing the median for spelling errors and QuillBot, the data displayed an even lower median of “percentage fakes” for the essays with spelling errors ($0.65 < 11.57$), indicating that essays with spelling errors are even harder to be detected by GPT-2 ODD. These findings can help answer the research question by not only identifying that there are differences between edited and unedited essays but also suggesting that GPT-2 ODD is more prone to certain post-editing techniques than others.

Discussion

After running the t-tests, the results strongly suggested that post-editing AI (ChatGPT Plus) generated essays do impact the result of the AI content detector (GPT-2 ODD). However, the impact of post-editing essays is not visible for all types of editing. From the five post-editing techniques used in this research, only two types of editing significantly impacted the accuracy of the AI content detector. The first type was introducing spelling mistakes in generated essays, and the second type was processing the essays with QuillBot.

Table 3 indicated that the median of the essays with spelling mistakes and essays processed with QuillBot was significantly lower than the unedited essays; this suggested that editing a generated essay with the above two techniques can trick the AI content detector into thinking those essays were written by humans. This is an issue because in addition to the problem of students using ChatGPT to cheat on exams (Cotton et al., 2023), they can also use different editing techniques to trick AI content detectors and fool the educator into thinking it is their own work. This finding is much similar to a concern mentioned in one existing study (Ventayen, 2023).

When further comparing the median of “percentage fakes” for essays with spelling errors and essays that were processed with QuillBot, the data showed a lower median for the essays with spelling errors. This result might indicate that AI content detectors such as GPT-2 ODD might make more incorrect detections if a student chooses to purposely introduce spelling mistakes into the generated essays.

Based on this research, it can be inferred that GPT-2 ODD exhibited significant performance variation, particularly in the context of edited essays. This finding aligned with the observations made in another study (Salminen et al., 2021).

Conclusion

This research discovered that GPT-2 ODD is sometimes inaccurate when it comes to detecting post-edited AI-generated essays, especially when the essays were edited with spelling mistakes or processed with QuillBot. These findings shed light on the research question by suggesting that human post-editing of AI-generated high school English essays will decrease the accuracy of AI content detectors. Furthermore, these findings should raise the attention of educators and AI content detector programmers.

Significance

ChatGPT has raised concerns about academic integrity (Cotton et al., 2023), and it was suggested that AI content detectors could be used to detect generated essays (Rodriguez et al., 2022). The findings from this research imply that AI content detectors are not always accurate when detecting generated essays, similar to the result from existing research (Rodriguez et al., 2022). In addition, the inaccuracy of detectors was amplified when the essays were edited, signaling that developers could improve the detection algorithm to adapt essays that students had edited.

The findings from this research are significant because it suggests that educators could not rely solely on current AI content detectors to detect students' generated essays, particularly if the student has edited the essay, where the accuracy of detecting such essays was previously unknown.

Although this study suggested the importance of improving the AI detection technology, it might not be the only way to solve the problem. Some studies suggested that educators can investigate ways to incorporate AI into their education instead of going against it (Popenici & Kerr, 2017). The homework students receive should not be able to be completed by simply using ChatGPT; instead, educators should create assessments that require students to demonstrate their critical thinking, problem-solving, and communication skills (Cotton et al., 2023).

Future research

To discover more potential limitations of AI content detectors, and to eliminate the limitations involved in this research, such as the lack of essay variations and lack of complexity in certain types of post-editing techniques, future research can experiment with more complicated essay editing techniques, or even try detecting different types of generated text, such as poems, limerick, et cetera, and not just limited to essays.

During this research, some other potential research questions surfaced and deserve more attention in further research. First, QuillBot appears to have multiple different editing modes, and the detectors' accuracy when detecting essays manipulated with different modes in QuillBot can be studied. Second, many different existing AI content detectors are available on the market, such as GPTZero and Turnitin, and their accuracy can also be studied. In addition, different large language models or versions of the GPT can be studied as well, as they might yield different results detected by AI content detectors.

Limitations

Type of Limitations

Outliers

During this research, it was surprising to see two outliers in the "percentage fake" results for unedited essays. P2 V1 only has 0.11% of "percentage fake" detected by GPT-2 ODD, and P5 V2 has 70.35%. Such variations in the detection of AI-generated essays are also observed in a similar study on GPT-2 Output Detector (Gao et al., 2022); however, that study does not contain any extreme values like 0.11%. These two outliers might indicate that AI content detectors can sometimes be inaccurate even if the generated essays were not edited. However, in this research where its goal is to evaluate the impact of post-editing on AI content detector, these outliers might be a limitation and negatively impact the accuracy of the experiment, because they cause the mean values to be skewed toward the outliers.

Prompt Variation

Another one of the limitations of the research is that prompt variation is limited. The random prompt selection process predominantly included former AP prompts, which are frequently used in AP classes. However, students who are not enrolled in AP Language and Composition courses may not encounter those AP prompts as frequently. As a result, the prompts employed in this study may not offer a comprehensive representation of the English prompts commonly used by high school students. Moreover, this study only incorporated five distinct versions of prompts, further limiting its scope. Consequently, the findings of this research may solely apply to specific types of English essay prompts. Future studies should consider examining a wider range of prompt variations, as certain essay types may exhibit differing levels of detectability.

Editing Techniques

One additional limitation of this research is that when replacing words with synonyms was used as an essay editing technique, only the third noun in each paragraph was changed to a synonym. Given that the introductory sentences of each paragraph in essays are often similar, it resulted in a significant number of the replaced third nouns being identical. As a result, there is a lack of variation in the nouns being edited, which might lead to the “percentage fake” detected for essays with synonyms unable to fully represent its population. With this limitation present in this research, future studies can implement more intricate essay editing techniques.

Validity

Although there are many limitations in this study, it does not undermine the validity of the data collected and observations made in this research. This is because all research processes carefully followed the method section; therefore, any potential human errors are minimized. The tools used in this research were also official programs that are available free for the public to use.

Acknowledgments

I would like to express my gratitude to all the high school English teachers at Singapore American School who provided the English prompts that were used in essay generation; they are the reason that this study could be completed. I would also like to thank my three student advisors in the Quest Program at Singapore American School for their unwavering support and encouragement throughout the entire research process.

References

- Aydın, Ö., & Karaarslan, E. (2023). Is ChatGPT Leading Generative AI? What is Beyond Expectations? *SSRN Electronic Journal*. <https://doi.org/10.2139/ssrn.4341500>
- Baidoo-Anu, D., & Owusu Ansah, L. (2023). Education in the Era of Generative Artificial Intelligence (AI): Understanding the Potential Benefits of ChatGPT in Promoting Teaching and Learning. *SSRN Electronic Journal*. <https://doi.org/10.2139/ssrn.4337484>
- Benzon, W. L. (2023). Discursive Competence in ChatGPT, Part 1: Talking with Dragons Version 2. *SSRN Electronic Journal*. <https://doi.org/10.2139/ssrn.4318832>
- Bommarito, J., Bommarito, M. J., Katz, J., & Katz, D. M. (2023). Gpt as Knowledge Worker: A Zero-Shot Evaluation of (AI)CPA Capabilities. *SSRN Electronic Journal*. <https://doi.org/10.2139/ssrn.4322372>
- ChatGPT*. (2023). Openai.com. <https://chat.openai.com/chat>
- Cotton, D., Cotton, P., & Shipway, J. (2023). Chatting and Cheating. Ensuring academic integrity in the era of ChatGPT. *EdArXiv Preprints*. <https://edarxiv.org/mrz8h/>
- Dehouche, N. (2021). Plagiarism in the age of massive Generative Pre-trained Transformers (GPT-3). *Ethics in Science and Environmental Politics*, 21, 17–23. <https://doi.org/10.3354/esep00195>
- Elkins, K., & Chun, J. (2020). Can GPT-3 pass a Writer’s turing test?. *Journal of Cultural Analytics*, 5(2).
- Floridi, L., & Chiriatti, M. (2020). GPT-3: Its Nature, Scope, Limits, and Consequences. *Minds and Machines*, 30(4), 681–694. <https://doi.org/10.1007/s11023-020-09548-1>
- Fröhling, L., & Zubiaga, A. (2021). Feature-based detection of automated language models: tackling GPT-2, GPT-3 and Grover. *PeerJ Computer Science*, 7, e443. <https://doi.org/10.7717/peerj-cs.443>

- Gao, C. A., Howard, F. M., Markov, N. S., Dyer, E. C., Ramesh, S., Luo, Y., & Pearson, A. T. (2022). *Comparing scientific abstracts generated by ChatGPT to original abstracts using an artificial intelligence output detector, plagiarism detector, and blinded human reviewers*. <https://doi.org/10.1101/2022.12.23.521610>
- Gozalo-Brizuela, R., & Garrido-Merchan, E. C. (2023). ChatGPT is not all you need. A State of the Art Review of large Generative AI models. *ArXiv.org*. <https://doi.org/10.48550/arXiv.2301.04655>
- Lavoie, A., & Krishnamoorthy, M. (2010). Algorithmic Detection of Computer Generated Text. *ArXiv.org*. <https://doi.org/10.48550/arXiv.1008.0706>
- OpenAI. (2022, November 30). *ChatGPT: Optimizing Language Models for Dialogue*. OpenAI; OpenAI. <https://openai.com/blog/chatgpt/>
- Popenici, S. A. D., & Kerr, S. (2017). Exploring the impact of artificial intelligence on teaching and learning in higher education. *Research and Practice in Technology Enhanced Learning*, 12(1). <https://doi.org/10.1186/s41039-017-0062-8>
- QuillBot. (2022). Quillbot.com; QuillBot. <https://quillbot.com/>
- Rawlinson, G. (1976). *University of Cambridge*. Cam.ac.uk. <https://www.mrc-cbu.cam.ac.uk/people/matt.davis/Cmabridge/rawlinson/>
- Rodriguez, J., Hay, T., Gros, D., Shamsi, Z., & Srinivasan, R. (2022). Cross-Domain Detection of GPT-2-Generated Technical Text. *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. <https://doi.org/10.18653/v1/2022.naacl-main.88>
- Salminen, J., Kandpal, C., Kamel, A. M., Jung, S., & Jansen, B. J. (2021). Creating and detecting fake reviews of online products. *Journal of Retailing and Consumer Services*, 64, 102771. <https://doi.org/10.1016/j.jretconser.2021.102771>
- Sumalinog, G. (2018). *COMMON GRAMMATICAL ERRORS OF THE HIGH SCHOOL STUDENTS: THE TEACHERS' PERSPECTIVE*. 5(10). <https://doi.org/10.5281/zenodo.1473359>
- Teo Susnjak. (2022). *ChatGPT: The End of Online Exam Integrity?* ArXiv; <https://www.semanticscholar.org/paper/ChatGPT%3A-The-End-of-Online-Exam-Integrity-Susnjak/8822357efe500caded16e603d21239be3a39547c>
- Thunström, A., Steingrímsson, S., & Gpt Generative Pretrained Transformer. (n.d.). *Can GPT-3 write an academic paper on itself, with minimal human input?* <https://hal.science/hal-03701250/document>
- Ventayen, R. J. M. (2023). OpenAI ChatGPT Generated Results: Similarity Index of Artificial Intelligence-Based Contents. *SSRN Electronic Journal*. <https://doi.org/10.2139/ssrn.4332664>
- Willems, J. (2023). ChatGPT at Universities – The Least of Our Concerns. *SSRN Electronic Journal*. <https://doi.org/10.2139/ssrn.4334162>
- Winter, J. de. (2023). *Can ChatGPT Pass High School Exams on English Language Comprehension?* Researchgate. Retrieved February 15, 2023, from https://www.researchgate.net/profile/Joost-De-Winter/publication/366659237_Can_ChatGPT_pass_high_school_exams_on_English_Language_Comprehension/links/63b9c3fcc3c99660ebd8847c/Can-ChatGPT-pass-high-school-exams-on-English-Language-Comprehension.pdf