Can We Trust ZeroGPT?

A Comprehensive Statistical Analysis of an AI-Generated Text Detection Tool

Edwin S. Hou

Los Gatos High School, CA

August 2024

Abstract

Recent developments in the generative AI field have revolutionized productivity, as tools like ChatGPT can respond to a variety of questions at the push of a button. Although convenient, the use and abuse of these tools come hand in hand. Menial tasks like sending emails have become easier than ever to automate, meanwhile academic integrity violations have become much more common. With students being able to produce entire human-sounding essays in a matter of seconds, improper regulation can have dangerous consequences for AI plagiarism in academia. This study aims to evaluate the reliability of ZeroGPT, a free AI-generated text detector tool trusted by millions of users. This paper explores a threshold sensitivity analysis to identify the optimal threshold that maximizes performance metrics such as precision, recall, and F1 score. Additionally, the paper also conducts a Bayesian analysis to determine the probability that a text was actually generated by ChatGPT given that ZeroGPT predicts it to be.

Can We Trust ZeroGPT?

A Comprehensive Statistical Analysis of an AI-Generated Text Detection Tool

Introduction

Background

In November of 2022, OpenAI's ChatGPT was released, showcasing the revolutionary leaps in AI technologies while also elevating AI into the public's awareness. As a chatbot using generative artificial intelligence to communicate with humans, ChatGPT was far beyond what other AI models promised, since previous technologies were relatively confined to research and had limited practical applications (ex. StockFish's AI chess engine). This breakthrough highlighted both the limitless potential for AI to automate mundane tasks and the possibility for individuals to commit plagiarism using AI-generated text, the latter garnering major attention and controversy. The launch of AI-generated text detection tools sparked an arms race of sorts, with OpenAI developers continuously improving their models to write more human-like text while detection software researchers work to develop more sophisticated methods to accurately differentiate AI-generated text from those written by humans. One such AI-generated text detection software that will also be the focus of this paper is ZeroGPT.

ZeroGPT

ZeroGPT is an AI detection tool designed to differentiate human written and AI text. From its website, ZeroGPT claims that it is a "Simple and Credible Open AI and Gemini detector tool for Free", and that "Millions of users trust ZeroGPT." (ZeroGPT) They claim that the software works by "a series of complex and deep algorithms [that] were developed by

ZeroGPT's team and they are backed by our in-house experiments and some highly reputable papers already published." (ZeroGPT) Given any piece of text, the tool would output one of the following results based on its predictions:

- Your text is Human written
- Your text is AI/GPT Generated
- Most of Your text is AI/GPT Generated
- Your text is Most Likely AI/GPT generated
- Your text is Likely generated by AI/GPT
- Your text contains mixed signals, with some parts generated by AI/GPT
- Your text is Likely Human written, and may include parts generated by AI/GPT
- Your text is Most Likely Human written, and may include parts generated by AI/GPT
- Your text is Most Likely Human written (ZeroGPT)

Additionally, ZeroGPT will also provide users with the total percentage of text suspected of being AI-generated alongside highlighting the sentences it believes to be AI-generated. The developers claim that the tool boasts a 98% accuracy rate with an error rate of less than 2% after analyzing over 10 million articles. This paper aims to evaluate the accuracy claims made by ZeroGPT.

Literature Review

With ChatGPT's release over a year ago, a considerable amount of research has been conducted on the topic of detecting generative AI. As academia takes interest in AI detection, several research papers have been published, including Aremu (2023) and Pegoraro et al. (2023).

Aremu's paper measured the performance of commonly used GPT detection tools that are

currently available on the market (Sapling AI, Crossplag, OpenAI's AI text classifier, ZeroGPT, GPTZero, and ContentAtScale AI detector). Their experiment consisted of passing various samples of either human-written or AI-generated essays through each AI detection tool and recording their results. Essay prompts included Gun Control, A Day at the Beach, The Benefits of Regular Exercise, and A Journey Towards Self-Discovery. The human-written essay samples were collected through crawling the web for texts written during the pre-GPT era (prior to 2022). Meanwhile, the AI-generated samples were produced by directly prompting ChatGPT to write essays on various topics. For this paper, we will focus on the experimental result concerning ZeroGPT. The results showed that ZeroGPT is largely effective at identifying human written texts with over an 80% accuracy rate. However, it performed poorly at reliably recognizing ChatGPT-generated essays, only achieving a 50% accuracy rate on average.

Pegoraro's paper provides a comprehensive assessment of the recent developments in ChatGPT detectors. Through analyzing prior literature published on the topic alongside researching the methods and principles behind GPT text detection, the paper can categorize AI detectors based on their functionality. The paper concludes that ZeroGPT "is specifically developed to detect OpenAI text but has limited capabilities with shorter text." (Pegoraro, 2023, p. 3)

Methodology

Data Collection

The input data for the experiment conducted in this paper is from the Kaggle dataset "AI vs Human Text", consisting of over 500,000 essays created by AI and written by humans. The

paper chose to utilize this common dataset rather than generating text directly from ChatGPT to reduce biases that may arise from ChatGPT remembering past conversations. ChatGPT being a tool tailored for holding conversations includes a feature to remember past dialogue with each user, allowing for a more personal experience. While this feature is beneficial in making ChatGPT sound more human-like, it becomes problematic in our use case as it creates the unaccounted variable of prior interactions. By using the Kaggle dataset, the experiment ensures that the input data is unbiased thereby eliminating the confounding variable of prior conversations and providing a more consistent set of AI-generated texts.

Experimental Procedure

The experiment consists of three major steps: Data Preparation, Setting Up Experimental Environment, and Extracting ZeroGPT Data Using Web Scraping.

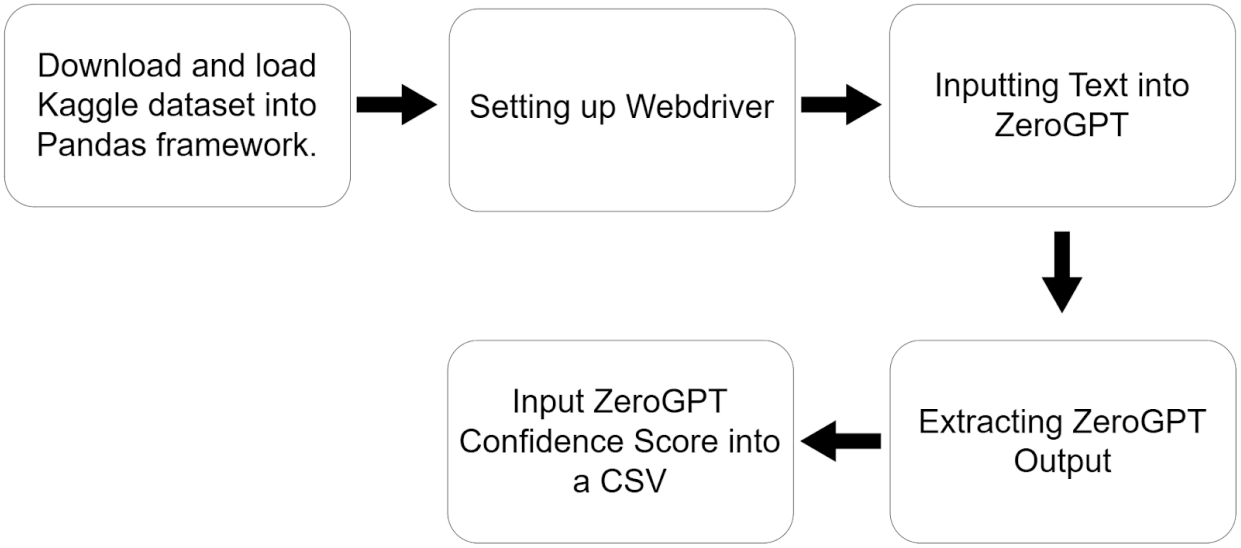

Fig. 1 Flow chart outlining the steps taken by the code conducting the experiment.

Data Preparation:

Download the AI_Human.csv from Kaggle

(https://www.kaggle.com/datasets/shanegerami/ai-vs-human-text). Load the data from the AI_Human.csv file into the Pandas Python framework in chunks of 100 rows to conserve CPU.

Setting Up Selenium Webdriver:

Start an instance of a Selenium web scraper in Python. Connect the web scraper to an existing Chrome tab using a debugger address to prevent bot detection from ZeroGPT.com.

Extracting ZeroGPT Data Using Web Scraping:

Input the text section of each row into ZeroGPT using selenium and click on the "Detect Text" button. Extract ZeroGPT's confidence score and highlighted sentences from the website and input them into an experimental result csv file. Repeat this step until sufficient data is gathered.

Experimental Results

As mentioned above, ZeroGPT's results are extracted and mapped to the ground truth (i.e., the actual identity of the text as either human-written or AI-generated) using an index system. The following, Fig. 2, is a snapshot of the result randomly selected from the result file. The complete experimental result csv file can be found on GitHub.

Text Id	zeroGPT Confidence	Is AI Generated?
36	0%	No
83	0%	No
1577	83.02%	Yes
1616	89.23%	Yes
1663	98.21%	Yes
3157	18.90%	No
3196	0%	No
3243	0%	No
4737	0%	No
4776	0%	No
4823	0%	No
6317	0%	No
6356	0%	No
6403	8.40%	No

Fig. 2, Result Set, randomly selected

From the input dataset of 500,000 records, a total of 48,900 records were used for the experiment. The resulting experimental CSV file consists of approximately 33% human-written texts with the remaining 67% being AI-generated. While initially seeming off, this percentage should not raise a source of concern because it accurately reflects the percentage of AI-generated text on the internet. A study conducted by researchers at the AWS AI Lab found that over 57.1 percent of all sentences on the internet were modified or generated by large language models (LLMs).

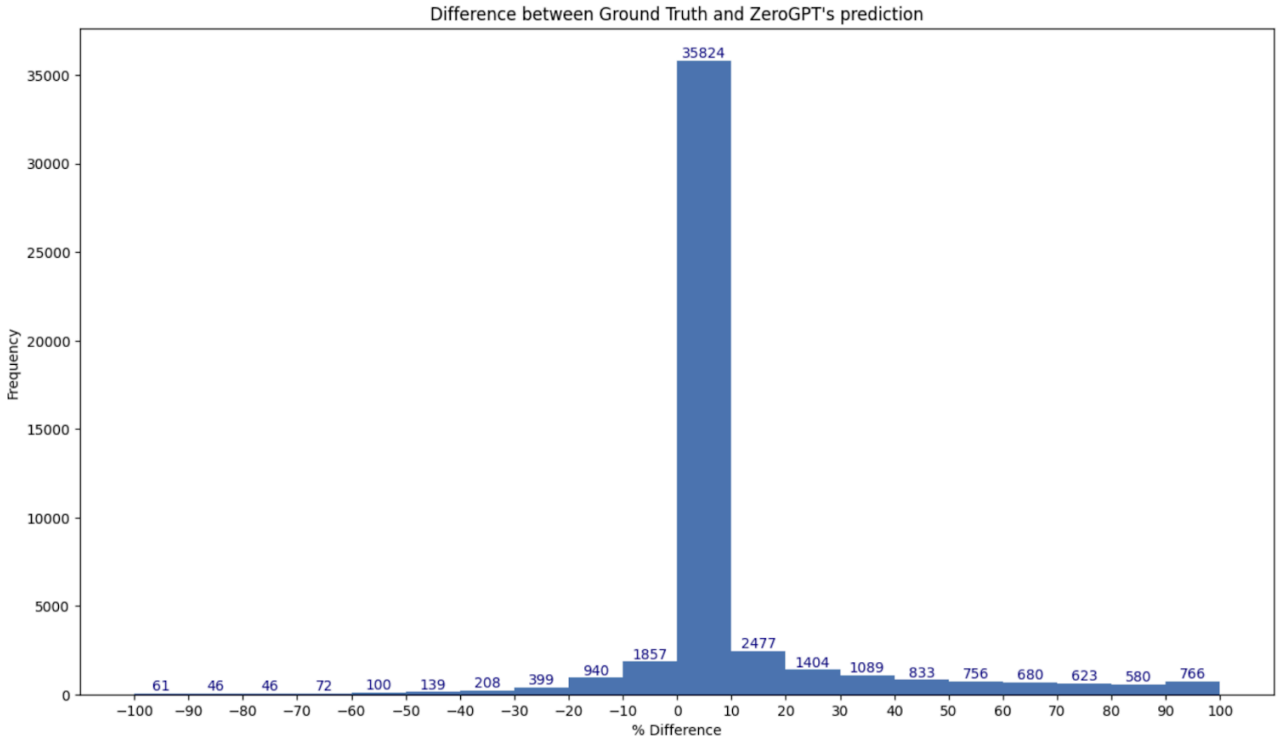

Fig. 3, Bar chart illustrating the difference between ZeroGPT's prediction and the ground truth.

The bar chart presented in Fig. 3 reveals that at a glance, ZeroGPT is fairly accurate in classifying a text's origin. However, additional statistical analysis is required to gain further insight into ZeroGPT's performance.

Statistical Analysis

Performance Metrics

Due to the binary nature of text being either human-written or AI-generated in addition to ZeroGPT returning a confidence value that lies in a range between 0% and 100%, a threshold sensitivity analysis allows for grouping of similar confidence scores into intervals. Further examination of ZeroGPT's performance on specific confidence intervals allows for a nuanced investigation of ZeroGPT's prevalence across different use cases. Moreover, conducting a Bayesian Statistical Analysis reveals the underlying false positive and false negative rates of ZeroGPT.

Threshold Sensitivity Analysis

A Threshold Sensitivity Analysis is conducted to determine the optimal threshold values for classifying text as either AI-generated or human-written.

In this paper, threshold refers to the largest ZeroGPT confidence score at which one determines a text to still be human-written. For example, when one operates with a threshold of 20%, they'll classify text with a ZeroGPT confidence score of 15% as human written.

Through examining performance at various threshold levels, one can identify the optimal threshold to maximize certain factors such as precision, recall, F1 score, and accuracy. The table below (Fig. 4) displays the key metrics (true positive, false positive, true negative, and false negative rates) across a variety of thresholds.

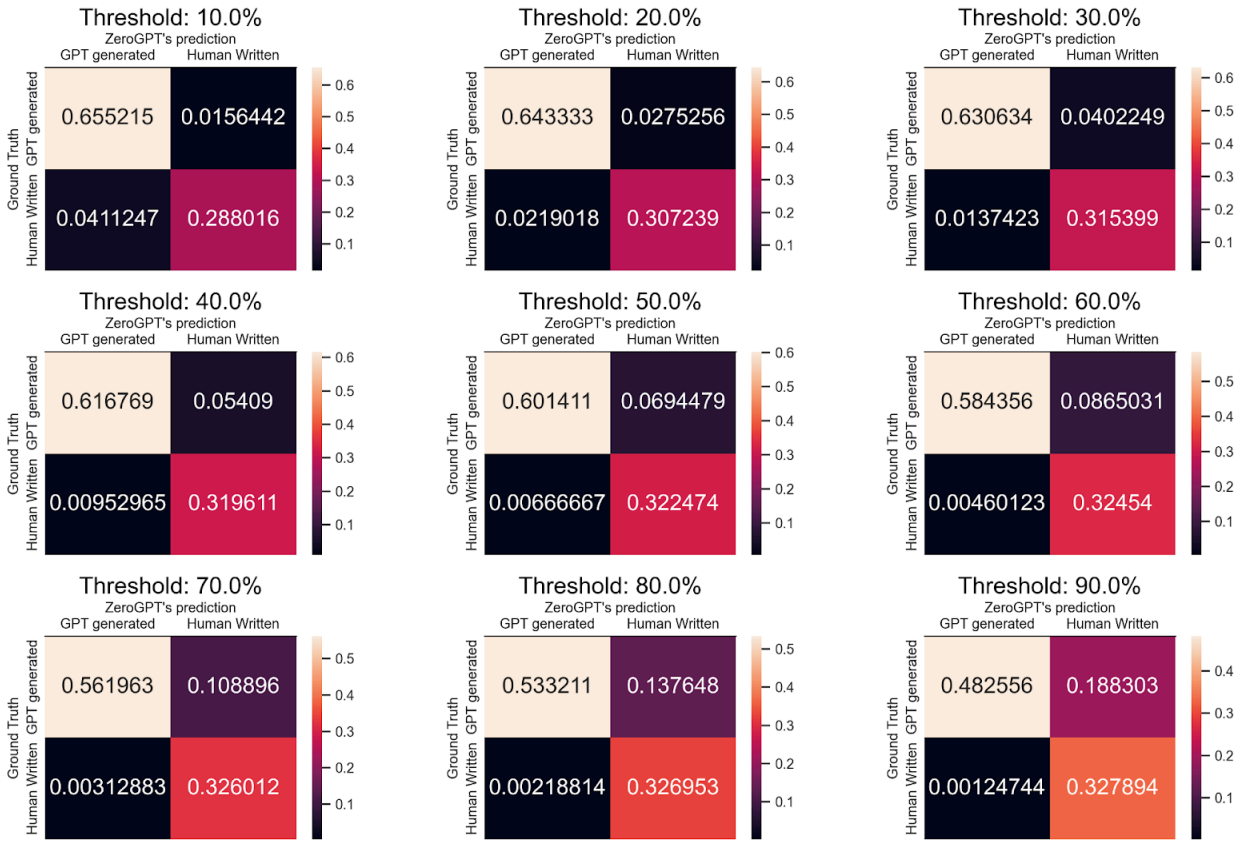

Fig. 4 Threshold Analysis: Key Metrics at Different Thresholds

Precision is defined as the ratio of true positive predictions to total positives predicted. It answers the question, "Of all the records that ZeroGPT labeled as positive, how many were actually positive?" A high value indicates that there are few false positives. The formula for precision is as follows:

$$Precision = \frac{True\ Positives\ (TP)}{True\ Positives\ (TP) + False\ Positives\ (FP)}$$

Recall is defined as the ratio of true positive predictions to the total actual positives. It answers the question, "Of all the actual positive records, how many did the ZeroGPT correctly identify?" A high recall indicates that there are few false negatives. Recall can be expressed as the following:

$$Recall = \frac{True\ Positives\ (TP)}{True\ Positives\ (TP) + False\ Negatives\ (FN)}$$

The F1 score is defined as the harmonic mean of the precision and recall scores. It provides a singular metric that balances both precision and recall, which is especially useful when the outcome of a binary event is uneven where one outcome occurs more frequently than the other. A high F1 score indicates that both precision and recall scores are fairly high. The formula for F1 score is the following:

$$F1\ Score = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}$$

Accuracy is defined as the ratio of correct predictions to the total number of predictions. Accuracy answers the question, "What proportion of all predictions did ZeroGPT accurately identify?" Accuracy can be calculated as the following:

$$Accuracy = \frac{True\ Positives\ (TP) + True\ Negatives\ (TN)}{Total\ Predictions\ Made\ (TP + TN + FP + FN)}$$

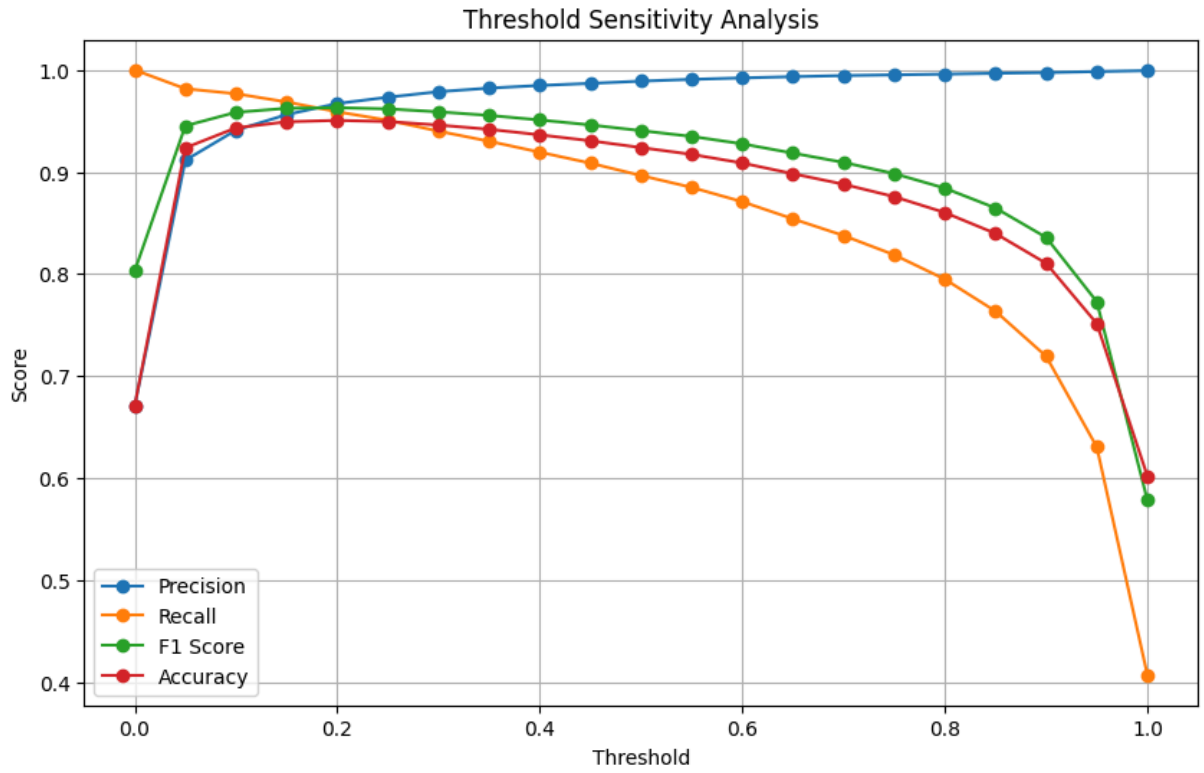

Fig. 5. Threshold Analysis: Performance Metrics at Different ZeroGPT Confidence Levels

From Fig. 5, it can be observed that:

Optimal Threshold: A threshold in the range of 0.20 to 0.30 seems optimal, providing a good balance between precision (around 0.97) and recall (around 0.95).

High Precision, Low Recall: Increasing the threshold beyond 0.50 leads to nearly perfect precision but at the cost of recall, reducing the tool's ability to identify all generated texts.

Low Thresholds: Very low thresholds (0.00 to 0.10) result in high recall but lower precision, suggesting the existence of many false positives.

Bayesian Statistical Analysis

Applying Bayes Theorem to the Threshold Sensitivity Analysis provides a better

understanding of the relationship between the ZeroGPT's confidence scores and the actual likelihood that a given text outside of the experimental dataset is AI-generated. The Bayes Theorem states the following: $P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$

A and B both represent certain events. P(A|B) on the left-hand side is the conditional probability of A occurring given that B occurred. On the right-hand side, P(B|A) is the probability of B occurring given that A occurred. P(A) and P(B) refer to the probability of events A and B occurring respectively. Bayes Theorem can be applied to this paper with event A representing a text actually being written by ChatGPT and event B representing a text that ZeroGPT predicted to be AI-generated. The probability P(A|B) measures the chances of a text being AI-generated given that ZeroGPT predicted it to be so.

The calculation used to determine the probability of a text being ChatGPT generated given that ZeroGPT predicts it to be is: $\frac{C \cdot True\ Positive\ (TP)}{C \cdot True\ Positive\ (TP) + (1-C) \cdot False\ Positive\ (FP)}$ where C denotes the probability of any given text being AI-generated. The numerator represents the probability that a text is generated by AI and ZeroGPT classifies it as such. The denominator represents the probability that ZeroGPT classifies any text as being AI-generated.

According to a Pew Research Center Survey, About 20% of U.S. teenagers have used ChatGPT to help with their schoolwork. Plugging in 0.2 for C across various thresholds yields the following table (Fig. 6):

THRESHOLD	P(A|B)
10%	79.9%
20%	88.0%
30%	92.0%
40%	94.2%
50%	95.8%
60%	96.9%
70%	97.8%
80%	98.4%
90%	99%

Fig. 6 Probability ZeroGPT correctly classifies a text as AI-generated across different thresholds given that 20% of all texts are AI-generated

Intelligent.com reports that approximately 30% of students during the 2022-2023 academic year used ChatGPT for schoolwork. Plugging in 0.3 for C across various thresholds yields the following table (Fig. 7):

THRESHOLD	P(A|B)
10%	87.2%
20%	92.6%
30%	95.2%
40%	96.5%
50%	97.5%
60%	98.2%
70%	98.7%
80%	99.1%
90%	99.4%

Fig. 7 Probability ZeroGPT correctly classifies a text as AI-generated across different thresholds given that 30% of all texts are AI-generated

The tables above (Fig. 6 and Fig. 7) show that the threshold percentage is directly proportional to ZeroGPT's accuracy rate. As such, increasing the threshold makes ZeroGPT's classification more accurate. However, after a certain point (approximately after 60%), further increasing the threshold yields diminishing returns in improved accuracy.

Discussion

Interpretation of Results

The extensive statistical analysis above suggests that ZeroGPT is only able to reliably classify a text's origin under specific circumstances. When using ZeroGPT with a low threshold, one runs into the issue of a high false positive rate. For example, assuming 20% of texts online are AI-generated and using a threshold of 10%, any given piece of text only has a 79.9% chance of being AI-generated when ZeroGPT claims it to be. In other words, there is a 20.1% chance that ZeroGPT falsely accused a text of being AI-generated. Having a one-in-five chance of a text being falsely flagged is significant enough to make the tool unreliable. However, when the threshold is increased to 60%, the false positive rate plummets to approximately 3.1% which makes the tool fairly reliable. Despite the benefit of decreased false positive rates, operating at a high threshold decreases the likelihood of detecting content that is actually plagiarized from AI. The largest decrease in false positive rate occurs when increasing the threshold from 10% to 30%. Increasing the threshold beyond 60% has diminishing returns in decreasing the false positive rate. From the Threshold Analysis, in general, it is recommended to use ZeroGPT with a 20% to 30% threshold to provide the best balance between a relatively low false positive rate and the ability to catch AI-plagiarized content.

Conclusion

Implications

As ChatGPT grows in popularity, it becomes necessary to develop a tool that can accurately detect AI-generated text to combat widespread AI plagiarism. It is only a matter of time before someone misuse the tool in fields such as academia where originality is of the utmost importance. While this paper initially set out to answer the black-and-white question of whether or not ZeroGPT is a reliable AI content detector, the conclusion reached is much more nuanced. It is determined that under specific circumstances, ZeroGPT could indeed accurately detect AI-generated content. As such, it is key to keep a text's context in mind before running it through ZeroGPT. The tool's authors' claim of it being able to accurately classify text 98% of the time is only true in certain scenarios. Below is a table (Fig. 8) illustrating the recommended threshold values across different contexts.

CATEGORY	RECOMMENDED THRESHOLD	REASONING
Students' Writing Homework	5-10%	Student homework is expected to be an authentic reflection of their learning and understanding. Even a small amount of AI-generated content can undermine the educational process. A very low threshold ensures that students are primarily responsible for their work, promoting learning and academic integrity.
Academic Contexts	10-15%	Academic integrity is paramount, and even small amounts of AI-generated content can compromise originality and authenticity. A lower threshold helps maintain high standards of originality.
Professional Contexts	20-30%	Professional documents often require precision and reliability. While some AI assistance may be acceptable for efficiency, ensuring the content remains primarily human-generated maintains trust and accountability.
Creative Contexts	30-40%	Creativity can be augmented by AI, but the originality of the creator is crucial. A moderate threshold allows for AI to assist while ensuring the core creative expression remains human.
Media and Journalism	10-20%	Accuracy and authenticity are critical in journalism. A lower threshold ensures that the information presented is reliable and primarily human-generated.
Marketing and Advertising	30-50%	Marketing content can benefit from AI tools for efficiency and creativity. A higher threshold is acceptable as long as the content remains engaging and meets ethical standards.
Personal Use	40-60%	Personal blogs and social media posts can have higher AI-generated content without major ethical concerns. The focus is more on expression and less on strict originality.

Fig.8, A table of recommended threshold values based on different scenarios.

Results from the table above (Fig. 8) in addition to the optimal threshold being between 20% and 30% suggest that ZeroGPT could be used to combat AI plagiarism within the context of professional, creative, and marketing use. Within these contexts, a false positive is less severe, allowing the benefits of catching AI-generated text to outweigh the risks. However, when it comes to judging students' homework or academia, it is not advised to take ZeroGPT's prediction into account due to the tool's high false positive rate. Even a false positive rate as low as 1 in 10, could result in a significant number of falsely flagged assignments, leading to unfair academic penalties and even expulsion. Overall, ZeroGPT's prediction should not be used as definitive proof. Instead, it should be used as initial suspicion to flag an offender about concerns over AI misuse. Other means of confirming AI-plagiarism such as comparing an offender's recently submitted text against their past writings should be used alongside ZeroGPT's results before reaching conclusions.

Limitations

One of the paper's limitations is the lack of variety in AI-content detectors. This paper only focuses on one detection tool, ZeroGPT, and as such, does not provide the full picture of AI-content detector development. Different AI-generated text detectors operate on different principles, and as such, this paper's results are not conclusive for GPT detectors in general and only pertains to ZeroGPT. Another limitation lies in the use of a subset of the available dataset. Applying the entire dataset would have made the results more generalizable and may reveal subtle patterns that are not apparent in a smaller sample. However, due to resource limitations, this paper only covers results with the use of the first 48,900 records.

Future Works

Future research may aim to address the aforementioned limitations. Comparing multiple AI-content detectors rather than focusing on only one which can provide a more comprehensive analysis of their developments as a whole. Doing so would also allow for comparison, which can offer valuable insight into their individual strengths and weaknesses.

Another avenue for continued research would be to conduct a semantic analysis of the sentences highlighted by ZeroGPT. This paper initially sought to determine certain writing cues that would tick off AI detectors and as such, collected the sentences ZeroGPT suspected of most likely being AI-generated when performing the experiment. Any future researchers interested in furthering this endeavor are free to use the highlighted sentences data already collected and stored at the GitHub Repository.

Furthermore, additional Bayesian Analysis could be conducted to determine the probability of a text being AI-generated despite ZeroGPT predicting it to not be. Such probabilities would provide a more comprehensive understanding of ZeroGPT's performance particularly in scenarios where false negatives have a significant impact. A high false negative rate would undermine the tool's purpose as a means of detecting AI-plagiarism.

References

(n.d.). AI Detector - Trusted AI Checker for ChatGPT, GPT4 & Gemini. Retrieved August 8, 2024 from https://www.zerogpt.com/

Morris, S., & Barrett, L. (2023, November 21). *Survey Reveals 1 in 5 US Teens Use ChatGPT for Schoolwork*. We Are Teachers. Retrieved August 8, 2024 from https://www.weareteachers.com/teenagers-use-chatgpt/

Nam, J. (2023, November 22). *56% of College Students Have Used AI on Assignments or Exams*. BestColleges.com. Retrieved August 8, 2024, from https://www.bestcolleges.com/research/most-college-students-have-used-ai-survey/

One-Third of College Students Used ChatGPT for Schoolwork During the 2022-23 Academic Year. (2023, September 5). Intelligent. Retrieved August 8, 2024, from https://www.intelligent.com/one-third-of-college-students-used-chatgpt-for-schoolwork-during-the-2022-23-academic-year/

Sidoti, O. (2023, November 16). *About 1 in 5 U.S. teens who've heard of ChatGPT have used it for schoolwork*. Pew Research Center. https://www.pewresearch.org/short-reads/2023/11/16/about-1-in-5-us-teens-whove-heard-of-chatgpt-have-used-it-for-schoolwork/

Aremu, T. (2023). Unlocking Pandora's box: Unveiling the elusive realm of ai text detection. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4470719 Detecting AI content in responses generated by CHATGPT, YouChat, and Chatsonic: The case of five ai content

detection tools. (2023). *Journal of Applied Learning & Teaching*, *6*(2). https://doi.org/10.37074/jalt.2023.6.2.12

Walters, W. H. (2023). The effectiveness of software designed to detect AI-generated writing: A comparison of 16 AI text detectors. *Open Information Science*, *7*(1). https://doi.org/10.1515/opis-2022-0158

Pegoraro, A. (2023, April 4). To ChatGPT, or not to ChatGPT: That is the question! *arXiv*, 6. arXiv.

Thompson, B. (2024, June 5). A Shocking Amount of the Web is Machine Translated: Insights from Multi-Way Parallelism. *AWS AI Labs*, 13.

Appendix

GitHub

Below is the link to the GitHub Repository that includes the source code that is referred to in this paper in addition to both the initial input and experimental datasets.

https://github.com/edwin-hou/FranklinResearch