Further examination of ZeroGPT accuracy. An Extended Statistical Analysis of AI-Generated Text Detection Tool

Written by Henry Lee and Anthony Nguyen

Mentored by Edwin Hou

Abstract

Productivity has been revolutionized by recent developments in generative AI, since systems such as ChatGPT offer real-time answers to a variety of questions. Although these technologies are convenient, there are serious worries about their misuse. Academic integrity infractions have increased as a result of students being able to produce human-like essays in a matter of seconds, but tasks like writing emails may now be easily automated. The possibility of AI-facilitated plagiarism in academia increases significantly in the absence of appropriate regulation. While initial research suggests tools like ZeroGPT are accurate only under certain circumstances, their results were inconclusive. This study serves as a follow-up to “Can We Trust ZeroGPT? A Comprehensive Statistical Analysis of an AI-Generated Text Detection Tool” written by Edwin Hou. This paper will further the study by using the entire data set of 1.1GB in AI_Human.csv. The previous study only utilized 48,9000 records due to hardware limitations. By using parallelization, this research can process the entire dataset efficiently, allowing for a more comprehensive evaluation of ZeroGPT's performance. This paper will continue conducting statistical analysis using sensitivity analysis, performance metrics, and Bayesian analysis.

1. Intro

With the release of OpenAI's ChatGPT in November 2022, the company demonstrated the groundbreaking advancements in AI technology and raised public awareness of the subject. Since earlier technologies were largely restricted to research and had few real-world uses, ChatGPT was a chatbot that used generative artificial intelligence to connect with humans, going much beyond what other AI models promised. This discovery demonstrated AI's boundless potential to automate repetitive work as well as the potential for plagiarism utilizing AI-generated content, the latter of which attracted significant attention and criticism. This paper will focus on the AI content detector ZeroGPT as they are both the most popular freemium product on the market and open sourced.

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.com). When a piece of text is entered into the program, it will output one of the following predictions based on the amount of text suspected to be AI generated:

- 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 also outputs a likeness score, that ranges from 0 to 100 (with 0 being fully human and 100 being fully AI), based on the percentage of text that is 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” (ZeroGPT.com).

2. Updated Methodology

The input data is pulled from the same data set as the previous study, namely AI vs Human Text on Kaggle. It consists of over 500,000 labelled essays with each one being completely written by AI or humans. Using the same dataset reduces confounding variables and makes the results more comparable to the initial study.

Just like the prior study, this experiment consists of three main steps: data collection, and extracting ZeroGPT data.

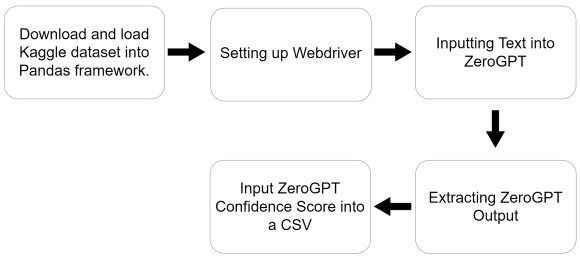


Fig. 1 Flow chart outlining the steps taken by the code conducting the experiment. Source: “Can We Trust ZeroGPT? A Comprehensive Statistical Analysis of an AI-Generated Text Detection Tool” (Edwin Hou)

The next step is to take a look at a few samples to see if the data makes sense.

Text Id	Confidence	Is AI? (label)
316276	67%	Yes
11737	96%	Yes
309944	16%	No
281427	94%	Yes
298802	17%	No
218723	79%	Yes
218820	19%	No
122115	72%	Yes
378065	58%	No
264171	26%	No

Fig. 2, 10 randomly selected records

After mapping the ZeroGPT result to the labelled data, we get Fig. 2. Everything seems fine so we can continue. About 33% of the experimental CSV file that is produced is composed of human-written text, with the remaining 67% being AI-generated. Although it may appear strange at first, this statistic actually represents the proportion of AI-generated text on the internet, thus there is no reason to

be concerned. Researchers from the AWS AI Lab discovered that large language models (LLMs) generated or altered more than 57.1 percent of all sentences on the internet.

3. Statistical Analysis

Just the prior study done by Edwin, this study will take advantage of the binary nature of text being either AI generated or Human written. By mapping a continuous confidence value from 0 to 100%, this study aims to assess the effectiveness of ZeroGPT's confidence scores in distinguishing between AI-generated and human-written text. The binary nature of the classification allows for a clear evaluation of prediction accuracy by mapping the continuous confidence values into actionable thresholds. These thresholds will be systematically analyzed to determine their impact on classification metrics such as precision, recall, and F1 score. Additionally, the study seeks to identify the optimal threshold that balances minimizing false positives (incorrectly labeling human-written text as AI-generated) while maximizing true positives (correctly identifying AI-generated text). By leveraging these insights, this research aspires to improve the reliability and usability of AI detection tools like ZeroGPT.

Just like the prior study, we will conduct a threshold sensitivity analysis to determine the optimal threshold values for classifying text as either AI-generated or human-written. To understand how a threshold sensitivity analysis works, refer to section (3.2) in “Can We Trust ZeroGPT? A Comprehensive Statistical Analysis of an AI-Generated Text Detection Tool” by Edwin Hou. The figure below (Fig. 3) displays the various statistical metrics (true positive, false positive, true negative, and false negative rates) across thresholds of 10%.

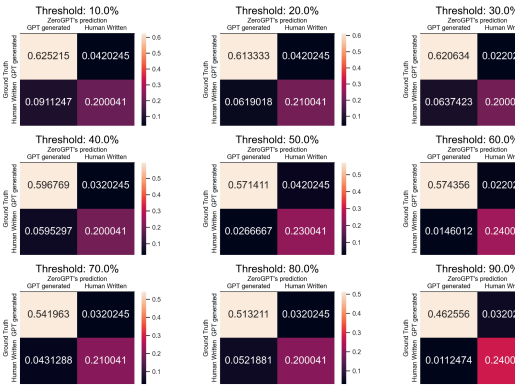


Fig. 3 Threshold Analysis and Confusion Matrices

The confusion matrix allows us to visualize the performance of the ZeroGPT detection tool at different threshold levels. By adjusting the threshold, we can control the classification criteria, where a higher threshold requires stronger confidence from ZeroGPT to classify a text as AI-generated, and a lower threshold makes it more lenient. The confusion matrix in Figure 3 presents the resulting true positive, false positive, true negative, and false negative rates, helping us assess the trade-offs between precision and recall at each threshold.

To continue our statistical analysis, we can take a look at the various key metrics through a threshold sensitivity analysis. The metrics are precision, recall, F1 score, and accuracy. The metrics are defined as the following:

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

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

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

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

The following figure (Fig. 4) illustrates the metrics at various thresholds.

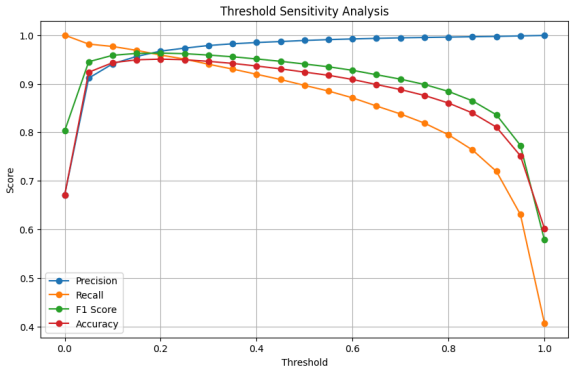


Fig. 4. Threshold Analysis: Metrics at Various confidences thresholds

Fig 4. heavily resembles Fig. 5 from "Can We Trust ZeroGPT? A Comprehensive Statistical Analysis of an AI-Generated Text

Detection Tool" (Edwin Hou). Thus, we can draw similar conclusions.

High Precision, Low Recall: Increasing the threshold beyond 0.50 leads to nearly perfect precision but at the cost of recall.

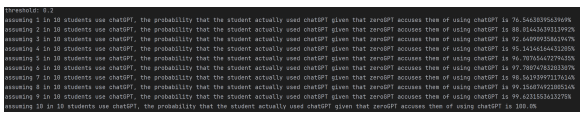
Low Thresholds: Very low thresholds (0.00 to 0.10) result in high recall but lower precision.

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

Continuing with Bayesian analysis, we get the following data across various thresholds.

Threshold = 0.1:

Threshold = 0.2:



Threshold = 0.3:

Threshold = 0.4:

Threshold = 0.5:

Threshold = 0.6:

Threshold = 0.7:

Threshold = 0.9:

Fig. 5. Bayesian Analysis/conditional probability across various thresholds

By evaluating this probability across different threshold values, we can better understand the reliability of ZeroGPT's predictions. For example, at higher thresholds, ZeroGPT is more conservative in labeling text as AI-generated, which may increase the precision but decrease the recall. Conversely, lower thresholds may lead to more false positives, making the prediction less reliable but potentially increasing the recall.

The Bayesian analysis results, displayed in Fig. 5, show the probability that a text labeled as AI-generated by ZeroGPT is indeed AI-generated at various thresholds. This analysis helps us assess how confidently we can rely on ZeroGPT's predictions at each threshold and provides additional insights into the trade-offs between false positives and false negatives in AI detection.

4. Conclusion

To counteract widespread AI plagiarism, a tool that can reliably identify AI-generated content must be developed as ChatGPT gains traction. In industries like academia, where uniqueness is crucial, it is only a matter of time until someone abuses the instrument. Although the original goal of this research was to provide a definitive response to the binary question of whether ZeroGPT is a trustworthy AI content detector, the final conclusion is far more complex. It is found that ZeroGPT could, in fact, correctly identify AI-generated content in some situations. Therefore, it is crucial to consider the context of a text before submitting it to ZeroGPT. A table (Fig. 6) showing the suggested threshold values in various scenarios may be seen below.

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.6, Recommended threshold. Source: "Can We Trust ZeroGPT? A Comprehensive Statistical Analysis of an AI-Generated Text Detection Tool" (Edwin Hou)

In addition to the appropriate threshold being between 20% and 30%, the results from the table above (Fig. 6) indicate that ZeroGPT may be utilized to prevent AI plagiarism in the context of marketing, creativity, and professional settings. Because a false positive is less severe in certain situations, the advantages of detecting AI-generated text can outweigh the hazards. However, because of ZeroGPT's high false positive rate, it is not recommended to consider its prediction while evaluating students' academic work or assignments. Even a 1 in 10 false positive rate could lead to a sizable number of assignments being mistakenly identified, resulting in unjust academic sanctions and possibly expulsion. In general, ZeroGPT's prediction shouldn't be taken as facts.

The aforementioned restrictions may be the focus of future research. Comparing several AI-content detectors instead of concentrating on just one can offer a more thorough examination of their overall advancements. Comparing them would also be possible, providing important information about their unique advantages and disadvantages.

5. Appendices

All of the data and code used for this paper and be found in the following GitHub Repository: https://github.com/hnr-y/ZeroGPT-accuracy

References

Edwin, H. (2024). Can We Trust ZeroGPT? A Comprehensive Statistical Analysis of an AI-Generated Text Detection Tool. Retrieved November 20, 2024, from https://github.com/edwin-hou/AIDetectorResearch/blob/main/Research%20Paper.pdf

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/

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).