IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

# AI-Generated Text Detection: A Review

[1]Dr. Ruchika Lalit, [2]Dr. Priyanka Bhutani, [3]Dr. Neha Verma, [4]Anshika jain
[1]Assistant Professor, [2]Assistant Professor, [3]Associate Professor, [4]PhD Scholar
[1]Department of Computer Science and Engineering,
[1]The NorthCap University, Gurugram, Haryana, India
[2]University School of Information, Communication & Technology,
[2]GGSIPU, Delhi, India
[3]IT Department,
[3]Member of Sustainable Development Goals (SDG), VIPS, GGSIPU, Delhi , India
[4]University School of Information, Communication & Technology,
[4]GGSIPU, Delhi, India

***Abstract:*** Large Language Models (LLMs) have made rapid strides in recent years, which has allowed them to excel at a wide range of activities including document completion and question responding. This has raised concerns about the unchecked usage of these models, which might lead to undesirable results like plagiarism, the creation of false news, spamming, etc. As LLMs must be used responsibly, accurate AI-generated content identification has become crucial. Several studies have attempted to solve this problem by including model signatures into text outputs or by watermarking texts with predetermined patterns. However, it has been observed that by a paraphrase attack, in which a light paraphraser is implemented on the LLM produced text. A range of AI detectors may be overwhelmed including those that employ watermarking methods, neural network-based detectors, and zero-shot classifiers. Furthermore, LLMs protected by watermarking methods are vulnerable to spoofing attacks, where a human adds covert watermarking signatures to human made text In this study, we examined the strengths and weaknesses of the watermarking and non- watermarking techniques for identification of text generated by AI. We looked at the soft watermarking technique for AI generated text and non-watermarked AI generated text and observed how it was vulnerable to spoofing and paraphrase attacks. This study reveals that the AI generated text detectors are unreliable under real-world conditions.

***Index Terms*** **- Large language models, AI generated content identification, DetectGpt, AI detectors.**

## I. Introduction

The rapid growth of artificial intelligence (AI) has resulted in the creation of sophisticated language models that can produce text that resembles that of a person[1]. Large language models (LLMs) have developed into extremely useful tools in a variety of fields, including information retrieval, content creation, and natural language processing. The prevalence of AI-generated text, however, raises questions about the veracity, integrity, and potential abuse of such material.

The ability to discern between language created by humans and that generated by LLMs is a crucial difficulty in maintaining the credibility of AI-derived content. In order to handle problems like plagiarism, the spread of false information, and content manipulation, this distinction is essential. Researchers have suggested many solutions, including watermarking and non-watermarking techniques, to address this issue and identify and confirm the source of text produced by AI.

The goal of watermarking techniques is to incorporate a digital watermark onto AI- generated text so that its source can be recognized and verified. The LLM's generated text is watermarked by the watermarking process by replacing a small portion of its parameters with a special watermark vector. On the other hand, non-

watermarking detectors rely on examining particular traits and statistical aspects of the text to distinguish between human-written and AI-generated content.

In this research paper, the strengths and weaknesses of watermarking and non-watermarking techniques for AI-generated text has been studied. Here, the robustness of Kirchenbauer et alsoft.'s watermarking technology against spoofing and paraphrase attacks is been observed. Also the evaluation of the scheme's capacity to safeguard AI-generated text's authenticity while being impervious to content manipulation has been examined.

Here, a keen investigation has been done on how paraphrase attacks affect non-watermarking detectors, such as trained classifiers and zero-shot classifiers and how well these detectors function when presented with professionally constructed paraphrased content. They use linguistic traits and statistical attributes to identify AI- generated text.

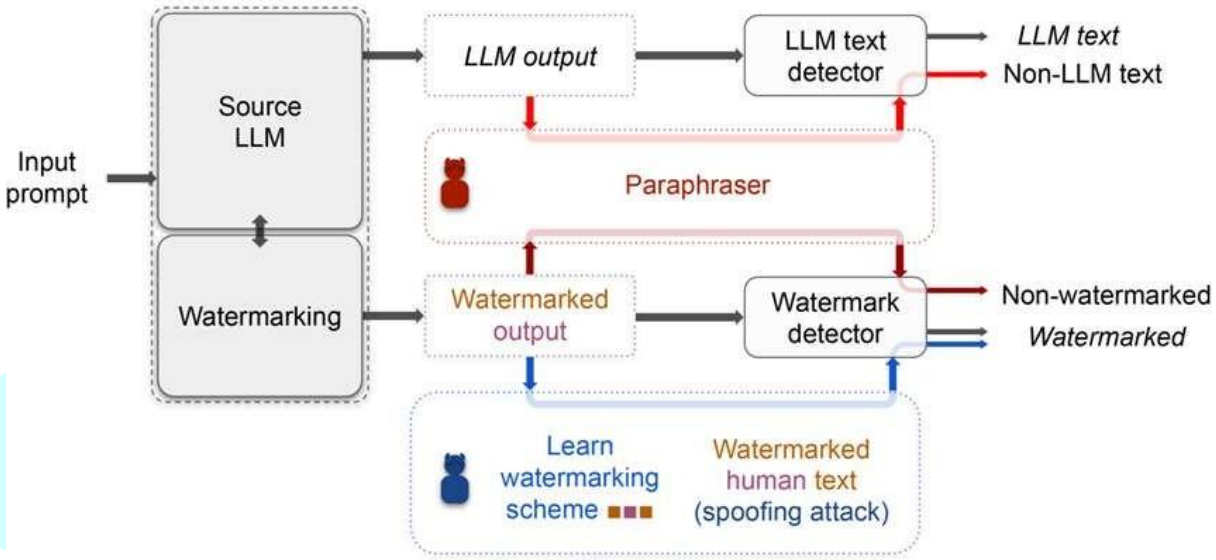


Figure 1: An Illustration of vulnerabilities of existing AI-detectors.[10]

It also provide insights into the flaws and restrictions of the current watermarking and non-watermarking techniques through in-depth tests and studies. We hope to open the door for the creation of stronger and more dependable methods for identifying and verifying AI- generated text by outlining the difficulties and potential flaws in these methods.

The findings of our study have significant implications for numerous fields that depend on AI-generated material, including journalism, content moderation, and intellectual property protection. The future work of our study is to create more reliable and durable AI systems by improving our understanding of the flaws and restrictions of current approaches.

## I. RESEARCH METHODOLOGY

### 2.1. Soft watermarking method for Ai-generated text

A technique for altering a small number of the model's parameters to apply a digital watermark to large language models (LLMs). A subset of the model's parameters are chosen, and they are then perturbed using a special watermark vector that is created using a secret key. The perturbation is intended to be both modest enough not to impair the performance of the model and large enough to be detectable.

Detection and embedding are the two stages of the watermarking process. The LLM's selected training parameters are subjected to the watermark vector during the embedding stage. As a result, a new set of parameters is created that contains the data from the watermark. The watermarked Data is used to generate text in the detection stage, and the text is then examined to find the watermark. Even when the LLM is optimized for a particular task or compressed to make it smaller, the watermark can still be seen.

A number of changes to the embedding process can be made to further increase the watermark's robustness, including picking the parameters to be watermarked at random and adding noise to the watermark vector. These changes aid in thwarting attempts to remove or alter the watermark.

The training method does not need to be altered or demand considerable computer resources when using the watermarking technique.

### 2.1.1. Paraphrasing attacks on watermarked Ai-generated text

In this study, we observed the susceptibility to paraphrase attacks of a watermarking technique suggested by Kirchenbauer et al. The purpose of the experiment was to determine whether effective paraphrasing could be used to remove the watermark signature from a large language model's (LLM) output while retaining the text's intended meaning.

Here, they have chosen the OPT-1.3B target AI text generator, a transformer-based model with a staggering 1.3 billion parameters, for the experiment. This LLM was chosen because it was well-trained to carry out text completion tasks on a huge corpus of data, giving it a suitable option for assessing the success of the watermarking technique.

It has used two paraphrasing models—one based on T5 and the other on PEGASUS—to carry out the paraphrasing attacks. While being smaller than the target LLM, these paraphrase models were optimized for paraphrasing tasks. This enabled us to replicate real-world scenarios where attackers could alter watermarked AI- generated text using smaller models. The experiment went like this: To test the watermarking scheme's capacity to maintain the integrity of the generated content, we first offered input prompts that either contained false information or fake news text. For these prompts, the target LLM produced watermarked results by hiding a signal in the text.

The LLM's watermarked outputs were then rephrased using the paraphrasing models. The language needed to be changed in order to remove the watermark signature while maintaining the intended meaning. We used a detection tool to assess the efficiency of the watermarking scheme and the impact of the paraphrase attacks. This method was designed to detect the presence or absence of a watermark in the output text. Both before and after applying the paraphrase attacks, we assessed the detection mechanism's accuracy.

The experiment's findings showed that paraphrase attacks were able to get around the watermarking system. The smaller paraphrasing models produced text that avoided being detected, which showed that the watermark signature had been removed. Significantly, the paraphrased language preserved the original intended meaning, demonstrating the watermarking system's potential susceptibility to paraphrasing attacks. These results highlight the need for more powerful and resilient watermarking approaches and highlight the shortcomings of the watermarking scheme in protecting against paraphrase attacks. The integrity and authenticity of AI-generated material must be guaranteed as it continues to play a large role across a variety of areas in order to stop the spread of false information and fake news.

### 2.1.2. Spoofing attacks on watermarked Ai-generated text

In this study, we discovered that watermarking schemes can be successfully faked, potentially resulting in the recognition of human-written texts as watermarked, raising questions about their dependability and efficacy. They adopted an attack strategy that entails memorizing the proxy green lists for the LLM's vocabulary's most commonly used words. Then they decrease the computational complexity while still generating useful results by choosing a smaller subset N (for example, N = 181) of frequently used phrases. In order to track pair- wise occurrences of these N words in the LLM outputs, it repeatedly query the watermarked LLM.

Here, it calculated the likelihood of a word arising given a prefix word using these observations. This probability score is used as a stand-in for calculating the prefix word's green list. It may modify the creation process to create texts that are recognized as watermarked and acquire insights into the watermarking patterns utilized by the LLM by learning these proxy green lists.

To recognize soft watermarked texts, we use the OPT-1.3B LLM [Zhang et al., 2022] as the target model. It compiled the green list scores for the prefix word "the" as an example to demonstrate how well our spoofing attacks work. It created an easy-to-use application that assists in creating meaningful passages by offering a list of prospective green list words that are categorized according to how well they perform at each phase. This tool makes it easier for users and hostile people to write paragraphs that are likely to be detected as watermarked.

By adopting the soft watermarking approach, the experimental results shows that even statements produced by hostile individuals may be watermarked with high confidence. This brings to light a major flaw in the scheme's ability to distinguish between documents created by LLMs and texts produced by humans. While the watermarking detector detects the presence of a watermark with accuracy, it is unable to determine whether a human writer or an Algorithm is responsible for creating the watermark pattern.

Both Watermarked and Non-Watermarked AI- Text detectors can be overwhelmed by using Spoofing and paraphrasing attack.

### 2.2. Non-Watermarked Ai-generated text

AI-generated texts are recognized by non- watermarking detectors without the use of overt watermarks. These detection techniques use different textual traits and statistical aspects to distinguish between content produced by humans and content produced by AI. Classifiers that have been trained and classifiers with zero shots are two examples of non-watermarking detectors.

Text datasets containing both human-written and AI-generated texts are used to train or fine- tune neural network-based models known as trained classifiers. These models learn to categorize incoming texts as either being created by humans or being created by artificial intelligence (AI) by analyzing patterns, linguistic traits, and statistical qualities included in the data. For instance, a popular trained classifier used for this purpose is OpenAI's RoBERTa-Large-Detector.

On the other hand, zero-shot classifiers make use of particular statistical characteristics specific to texts produced by AI in order to identify their source. These classifiers are trained using datasets that include both human- written texts and a sample collection of texts produced by artificial intelligence. They can categories unknown documents as either human- or AI-written by using these statistical patterns. The zero-shot classifiers proposed by Mitchell, Gehrmann, Ippolito, and Solaiman are notable examples.

#### 2.2.1. Paraphrasing attacks on non-watermarked Ai-generated text

Here, we studied the vulnerability of trained classifiers and zero-shot classifiers, two types of non-watermarking detectors, to a paraphrase attack. These detectors base their detection on finding particular patterns that are peculiar to texts produced by AI. Yet, the tests show that they are susceptible to our attack methodologies.

they used a T5-based paraphrasing model with 222M parameters to change the output texts produced by the GPT-2 model, and a pre- trained GPT-2 Medium model with 355M parameters to assess the efficiency of our attack. 200 sections from the XSum dataset were used in the attack [Narayan et al., 2018].

The effectiveness of the non-watermarking detectors is greatly decreased by the paraphrase attack, according to the results. For instance, DetectGPT, a trained neural network-based detector, saw a decline in AUROC scores from 96.5% to 59.8% as a result of our attack. The performance of the zero-shot detectors put forth by Solaiman et al., Gehrmann et al., and Ippolito et al. was equally subpar after our DetectGPT, a trained neural network-based detector, saw a decline in AUROC scores from 96.5% to 59.8% as a result of our attack. The performance of the zero-shot detectors put forth by Solaiman et al., Gehrmann et al., and Ippolito et al. was equally subpar after our attack. Although trained neural network-based detectors outperformed zero-shot detectors, such OpenAI's RoBERTa-Large-Detector, they were still vulnerable to our attack. For instance, even with a realistic false positive rate of 1%, the genuine positive rate of the RoBERTa- Large-Detector dropped from 100% to about 80% following our attack. A threat could further reduce the true positive rate to 60% by sending numerous inquiries to the detector. It is clear that the text has been effectively paraphrased to avoid being recognized by the trained classifiers while still maintaining coherence and meaning.

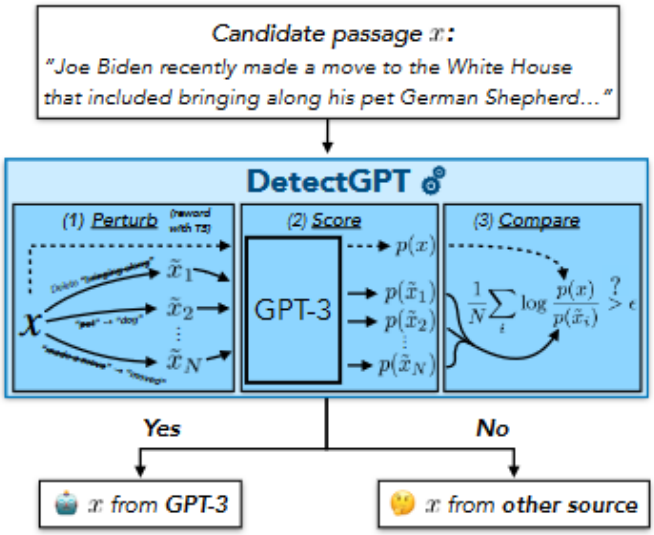


Figure 2: sample model used by DetectGpt[4]

These non-watermarking detectors are essential for detecting AI-generated texts in a variety of applications, including the detection of disinformation, content moderation, and plagiarism. Yet, as shown by the paraphrase attack, The findings highlight the shortcomings of non-watermarking detectors in accurately detecting AI-generated texts. Therefore, there is a critical need for enhanced detection methods that can withstand such attacks and consistently discriminate between writings produced by AI and those written by humans.

## II. DISCUSSIONS

The methods presented in this research paper shed light on the challenges and vulnerabilities associated with watermarking and non-watermarking techniques for detecting AI-generated text. In the context of watermarking, the study demonstrates the feasibility of paraphrase attacks that can effectively remove watermarks from AI-generated content while preserving the intended meaning. This highlights the need for more robust watermarking approaches to protect against such attacks, especially as AI-generated content continues to play a significant role in various domains where the spread of false information is a concern.

Furthermore, the research unveils the shortcomings of non-watermarking detectors, both trained and zero-shot classifiers, when faced with paraphrase attacks. These detectors rely on specific patterns to distinguish between human and AI-generated text, but the study shows that these patterns can be effectively manipulated, reducing their accuracy. This underscores the importance of enhancing detection methods that can withstand sophisticated attacks and consistently differentiate between AI-generated and human-written text.

In summary, the findings emphasize the critical need for advancements in both watermarking and non-watermarking techniques to ensure the integrity and authenticity of AI-generated content. As AI-generated text becomes increasingly prevalent, addressing these vulnerabilities is crucial for applications such as disinformation detection, content moderation, and plagiarism prevention. Future research should focus on developing more resilient and adaptive methods to stay ahead of evolving attack strategies in the realm of AI-generated text.

## III. CONCLUSION

In this study, it was observed that the strengths and weaknesses of watermarking and non-watermarking techniques for text generated by AI. Here, the soft watermarking technique suggested by Kirchenbauer et al. was observed. Also, how it was vulnerable to spoofing and paraphrase attacks.

We observed paraphrase attacks can be used to get around the watermarking approach, which inserts a digital watermark by changing a small portion of the model's parameters. Here, we have seen that they were able to successfully erase the watermark signature while maintaining the text's intended meaning by using smaller paraphrase models. This illustrates how ineffective the watermarking system is at blocking content modification.

In addition, we observed that they proved that spoofing attacks on text produced by watermarked AI were effective. It was able to create texts that were recognized as watermarked even when they were produced by antagonistic humans by learning proxy green lists and modifying the generation method. This raises questions regarding the watermarking scheme's dependability and capability to distinguish between human-written and AI-generated texts.

We also observed how susceptible non- watermarking detectors were to attacks that involved paraphrase, including trained classifiers and zero-shot classifiers. Their tests demonstrated that paraphrasing can seriously weaken these detectors, which rely on spotting particular patterns specific to texts produced by artificial intelligence. Improved detection methods are essential, as even the finest detectors showed a significant decline in effectiveness.

The results of this study shed light on the shortcomings of current watermarking and non- watermarking techniques in thwarting attacks on text produced by AI. In order to counteract false information and preserve trust, it is becoming more and more important to ensure the integrity, validity, and detection of AI- generated material.

Both Watermarked and Non-Watermarked AI-Text Detection methods can be overwhelmed by using Spoofing and paraphrasing attacks.

It is crucial to create stronger, more resilient watermarking systems going future so they can fend off spoofing and paraphrase attacks. Moreover, improvements in non-watermarking identification techniques should be sought to boost their dependability and efficiency in correctly detecting texts produced by main AI. In the end, this research offers insightful information for researchers, practitioners, and policymakers alike about the

difficulties and prospects in safeguarding AI-generated text. By solving these flaws, we may advance the creation of reliable AI systems and reduce the dangers posed by AI-generated content.

.

## REFERENCES


**[1]** J. Kirchenbauer, J. Geiping, Y. Wen, J. Katz, I. Miers, and T. Goldstein, "A Watermark for Large Language Models," in *Proceedings of the 40th International Conference on Machine Learning*, PMLR, Jul. 2023, pp. 17061–17084. Accessed: Sep. 15, 2023. [Online]. Available: https://proceedings.mlr.press/v202/kirchenbauer23a.html

**[2]** PEGASUS-based paraphrasing-https://huggingface.co/tuner007/pegasus_paraphras

**[3]** Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Stoyanov, V. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach

**[4]** Mitchell, E., Lee, Y., Khazatsky, A., Manning, C. D., & Finn, C. (2023). DetectGPT: Zero-Shot Machine-Generated Text Detection using Probability Curvature.

**[5]** OpenAI. Ai text classifier. https://platform.openai.com/ai-text-classifier,2023.

**[6]** Shalini Ghosh, Oriol Vinyals, Brian Strope, Scott Roy, Tom Dean, and Larry Heck. 2016. Contextual LSTM (CLSTM) models for large scale NLP tasks. CoRR,abs/1602.06291.

**[7]** Sichel, H. S. (1975). On a distribution law for word frequencies. *Journal of the American Statistical Association*, *70*(351a), 542-547.

**[8]** Narayan, S., Cohen, S., &Lapata, M. (2018, November). Don't Give Me the Details, Just the Summary! Topic-Aware Convolutional Neural Networks for Extreme Summarization. In *2018 Conference on Empirical Methods in Natural Language Processing* (pp. 1797-1807). Association for Computational Linguistics.

**[9]** Baayen R.H.: Word Frequency Distributions. Kluwer Academic Publishers, Amsterdam,The Netherlands,(2001)

**[10]** Vinu Sankar Sadasivan, Aounon Kumar, Sriram Balasubramanian, Wenxiao Wang, Soheil Feizi*:* Can AI-Generated Text be Reliably Detected? CoRR abs/2303.11156

**[11]** OPT -https://huggingface.co/facebook/opt-1.3b.