AJCSE
ASIAN JOURNAL OF COMPUTER SCIENCE ENGINEERING

ISSN 2581 – 3781

Available Online at www.ajcse.info
Asian Journal of Computer Science Engineering 2025;10(2):1-13

RESEARCH ARTICLE

Robustness Analysis of Maryland Watermark to Paraphrasing Attacks Under Advanced Techniques

Rajiv Kumar
Director - Data Science
Oracle USA
groverrajiv1984@gmail.com

Received on: 06/04/2025; Revised on: 07/05/2025; Accepted on: 08/06/2025

***Abstract*—Developing trustworthy techniques to recognize and validate machine-generated material is necessary due to the proliferation of large language models (LLMs) and the possibility of abuse. The Maryland Watermark proposed is a notable technique that embeds identifiable signatures into text generated by LLMs. This study investigates the robustness of the Maryland Watermark against paraphrasing-based evasion strategies in AI-generated text. With growing concerns over detecting machine-generated content, watermarking methods like Maryland, which subtly alter token selection probabilities, are critical for content attribution. Using the Mistral-7B-Instruct-v0.2 model and prompts from the DAIGT dataset, 1,000 documents (500 watermarked) were generated and subjected to three types of attacks: paragraph-based paraphrasing using a Seq2Seq model trained on kPar3, sentence-level paraphrasing using a T5-based ChatGPT Paraphraser, and word-level synonym substitution using a POS-aware WordNet approach. Evaluation metrics included watermark detectability (z-score, TPR, FPR), semantic similarity, and text quality (perplexity). Results show that paragraph-based paraphrasing yielded the lowest perplexity (19.53) while degrading semantic similarity most significantly, followed by sentence-based paraphrasing (perplexity 24.89). Recursive paraphrasing reduced watermark detection initially but showed recovery in detection accuracy in subsequent iterations. Word replacement attacks achieved high TPRs (95.78% for noun substitution and 39.76% for 25% token replacement), indicating their ineffectiveness. Overall, the Maryland Watermark remains robust against word-level modifications but is moderately vulnerable to advanced paraphrasing that alters semantic integrity.**

***Keywords*—Maryland Watermark, Large Language Models (LLMs), watermark robustness, paraphrasing attacks, watermark removal, detection accuracy, text authentication.**

I. INTRODUCTION

The capacity to distinguish between text produced by machines and text authored by humans is the foundation of several strategies to lessen the possible risks associated with generative language models. This includes well-known negative effects like models being often used maliciously for things like social media bots, phoney product evaluations, Wikipedia content production, or the automatic creation of focused spear phishing assaults against susceptible groups [1]. Furthermore, the capacity to monitor and record machine-generated text usage may mitigate the negative effects of future issues that have not yet been noticed. These issues could include anything from the overabundance of blogs and other online content produced by LLM to the contamination of upcoming training data [2][3]. It can be challenging to identify machine-generated writing, unfortunately. With LLMs becoming more widespread, machine-generated text might become a major source of spam, social media bots, and useless information on the internet [4][5]. By making it possible to identify and record LLM-generated text, watermarking is an easy and efficient way to lessen these damages. In January 2023, the first Large Language Model (LLM)-focused watermarking technique, known as the Maryland Watermark, was introduced [6][7]. This technique biases textual generation toward specific words, enabling the detection of watermarked content. Further developments have expanded on this idea, incorporating methods such as embedding rare Unicode characters within the generated text. The primary purpose of watermarking is to assert copyright and verify content authenticity [8]. Unlike traditional watermarks that are confined to a single media type, multimodal watermarks can be embedded across various formats, enhancing security and flexibility. A straightforward and efficient method for reducing these damages is watermarking, which makes it possible to identify and record LLM-generated text [9].

A key concern in watermarking is its robustness against manipulation. Effective watermarks should remain intact even when content is transformed, compressed, or paraphrased. This is particularly important in digital rights management, piracy detection, and verifying content authenticity [10]. However, sophisticated paraphrasing techniques, including adversarial attacks, threaten the integrity of watermarks by altering textual content while preserving its meaning [11][12]. Watermarking in paraphrasing attacks refers to the technique of embedding hidden, traceable patterns within text generated by "Large Language Models" (LLMs) to identify and verify the origin of content, even after it has been rephrased. As paraphrasing attacks involve rewording or restructuring original text to evade detection tools and conceal plagiarism, traditional watermarking methods often fail. To address this, advanced watermarking strategies are designed to be resilient against such attacks by encoding semantic or syntactic signals that remain intact despite rephrasing. These robust watermarks help in preserving content integrity, enabling authorship verification, and preventing the misuse of AI-

generated text in academic, journalistic, and content-driven domains [13][14]. In order to determine if watermarked text can survive changes without losing its detectability, this study evaluates the Maryland Watermark's resilience to paraphrase attacks. The findings will provide insights into watermark durability and its effectiveness in combating AI-generated misinformation, including its potential role in mitigating phishing threats.

A. Motivation and significance of this study

Growing concerns about the integrity and detectability of AI-generated information, particularly in light of adversarial paraphrase strategies meant to avoid watermark detection, are the driving force behind this study. As AI-generated texts become increasingly prevalent, ensuring the robustness of watermarking methods like the Maryland Watermark is critical for accountability, copyright enforcement, and trust in digital content. This research is significant as it not only evaluates the watermark's resilience against realistic and varied paraphrasing attacks but also provides practical insights into its limitations and strengths, guiding future developments in secure and tamper-resistant watermarking systems. The following summarization of this paper are:

- Using a variety of models and datasets, the study offers a methodical, multi-stage approach to assess the Maryland Watermark's resilience against actual cyber-attacks.
- It introduces and compares three distinct types of paraphrasing attacks—paragraph-level, sentence-level, and word-level—demonstrating their varying impacts on watermark detectability, text quality, and semantic fidelity.
- The study explores the effect of recursive paraphrasing, revealing that while initial iterations degrade detection performance, further iterations lead to partial recovery, indicating diminishing returns in such evasion strategies.
- Detailed performance metrics (z-score, TPR, FPR, cosine similarity, and perplexity) provide empirical evidence on how each attack influences watermark robustness, fluency, and semantic coherence.
- The results validate that the Maryland Watermark is highly resilient to word-level substitutions but moderately vulnerable to sophisticated paraphrasing, offering insights into the trade-off between watermark durability and semantic preservation.

The following paper are organized as: Section II provide the literature review, methodology with each step discussed in Section III, results and discussion of given methodology evaluated in Section IV, Conclusion and future work discussed in Section V.

II. Literature Reviews

In light of growing worries about the proliferation of language model (LM)-generated content on the internet, watermarking is viewed as a morally sound method of verifying if a text was produced by a model. A signal that can be later identified is included in the generated output by a number of modern watermarking approaches that marginally alter the output probabilities of LMs.

Rastogi and Pruthi (2024) Concerns regarding the early text watermarking ideas' resilience to paraphrase have been widely debated. Some strategies are purposefully created these days and are said to be resistant to paraphrase. Such watermarking systems, however, fail to sufficiently take into consideration how easily they can be reverse-engineered. They demonstrate that by gaining access to a restricted set of generations from a black-box watermarked model, it can significantly boost the efficacy of paraphrase attacks to avoid watermark detection, making the watermark useless [15].

Zhang et al. (2024) conduct a thorough analysis of the state-of-the-art LLM watermark scheme's susceptibility to a new green list stealing attack. A mixed integer programming issue with limitations is how they formulate the attack. They test their attack under a full threat model, which includes an extreme case in which the attacker is completely ignorant, does not have access to the watermark detector API, and is unaware of the watermark injection/detection technique or the LLM's parameter settings. Long-term tests on LLMs, including OPT and LLaMA, show that their attack can effectively remove the watermark and steal the green list in any situation [16].

Barman et al. (2024) claim that the picture watermarking techniques used today are brittle and vulnerable to visual paraphrasing assaults. There are two stages to the suggested visual paraphraser's operation. First, it uses KOSMOS-2, one of the newest and most advanced image captioning systems, to create a caption for the provided image. The final image has no watermarks and is a visual paraphrase. practical results show that watermarks can be successfully removed from photos via visual paraphrase attacks. This study offers a critical evaluation, empirically demonstrating how susceptible current watermarking methods are to visual paraphrase attacks [17].

Hongyan Chang et al. (2024) present the smoothing attack and demonstrate how vulnerable current statistical watermarking techniques are to small text changes. Specifically, an adversary can smooth out the distribution disruption induced by watermarks using a weaker language model. While avoiding the watermark detector, the generated text that is produced is of a quality that is comparable to that of the original (unwatermarked) model. Their attack exposes a basic flaw in a variety of watermarking methods [18].

Idrissi et al. (2023) main objective was to ensure the ethical use of LLMs in AI-driven text synthesis by developing a novel methodology for the detection of synthetic text. The study begins by reproducing results from an earlier baseline study, highlighting its vulnerability to changes in the underlying generation model. They next suggest a novel watermarking strategy and rigorously test it using generated text that has been paraphrased to see how resistant it is. Results from experiments demonstrate how reliable their suggestion is in comparison to the watermarking technique [19].

Table I highlights how many watermarking methods are vulnerable to paraphrasing attacks. Simple rewording or visual transformations can effectively remove watermarks, even without knowledge of the watermarking system. This shows the need for more robust and paraphrase-resistant watermarking techniques.

TABLE I. Watermark Against Paraphrasing Attacks using various LLM techniques

Author(s)	Year	Focus Area	Technique/Attack Proposed	Key Contribution	Vulnerability/Challenge Addressed
Rastogi and Pruthi	2024	Text Watermarking	Paraphrasing Attack	Demonstrated that paraphrasing can effectively evade detection with few examples	Black-box paraphrasing can nullify watermarks.
Zhang et al.	2024	Text Watermarking	Green List Stealing Attack (Mixed Integer Programming)	Propose a green list stealing attack even under extreme threat models	Effective even without access to watermark scheme or detector API
Barman et al.	2024	Image Watermarking	Visual Paraphrase Attack using KOSMOS-2	Showed that generating captions removes watermarks from images	Visual paraphrasing renders image watermarking ineffective
Hongyan Chang et al.	2024	Text Watermarking	Smoothing Attack using weaker language model	Smoothing distribution perturbed by watermark avoids detection	Even minor edits bypass statistical watermark detectors
Idrissi et al.	2023	Synthetic Text Detection and Watermarking	New watermarking method + robustness evaluation (paraphrased)	Developed robust watermarking approach validated against paraphrased text	Baselines are fragile under different LLM generations

III. Methodology

This methodology covers text generation, watermarking, and removal attacks like paraphrasing and word replacement. It evaluates watermark robustness using detection accuracy, document similarity, and perplexity analysis to assess resistance against manipulation.

A. Text Generation

This section will go over the concepts behind providing text to models, as well as how the model chooses the next word in a sentence.

1) Tokenization

Tokenization structures textual data into tokens, enabling efficient language model training. It breaks text into words, subwords, or characters, enhancing model performance in tasks like text generation, classification, and translation.

- Token: A string of characters that frequently denotes a word or subword is called a token.
- Vocabulary: A vocabulary is the collection of tokens recognized by a tokenization function [20].

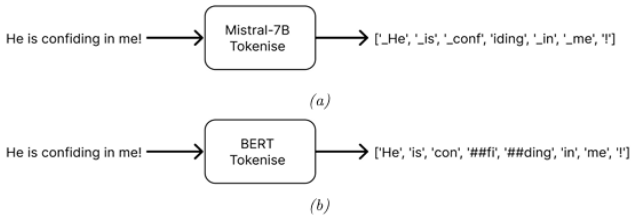


Fig. 1. Tokenization Function Comparison

The training of tokenization models creates a unique vocabulary based on the text corpus, as seen in Figure 1. This uniqueness carries through the transformer-based architecture, impacting watermarking in language models.

2) Generation Strategies

The Causal Generation is a text generation task that predicts the next token by examining previous tokens. It covers model outputs and generation techniques. Causal Language Models, like ChatGPT, produces a collection of values, known as logits. From these logits, I obtain a probability distribution for the next token with the application of the Softmax function. The created probability distribution is used for multiple generative strategies, including greedy generation, multinomial sampling, and top-p sampling [21]. Each generation strategy serves a distinct purpose. Greedy generation selects the highest-probability token, ensuring deterministic output. Multinomial sampling introduces randomness by selecting tokens based on their probability distribution. Top-p sampling refines this by limiting selection to the most probable tokens, ensuring diversity while reducing the likelihood of unusual tokens.

B. Generating Watermarks

A watermark is a faint figure or signature designed to represent ownership or authorship.

1) Maryland Watermark

The author [22] proposes a new method of watermarking text that takes advantage of the probabilistic nature of text generation. This section will go over the ideas of the Maryland Watermark, a technique that laid the foundations of LLM-focused watermarking.

There are two primary watermarking techniques from [22], Hard Watermarking and Soft Watermarking. Hard Watermarking only allows the causal model to select tokens from the green list, whereas Soft Watermarking chooses to increase the probability of the green tokens within the probability distribution. Figure 2 outlines the technique and the change that the soft-watermarking process has on the probability distribution.

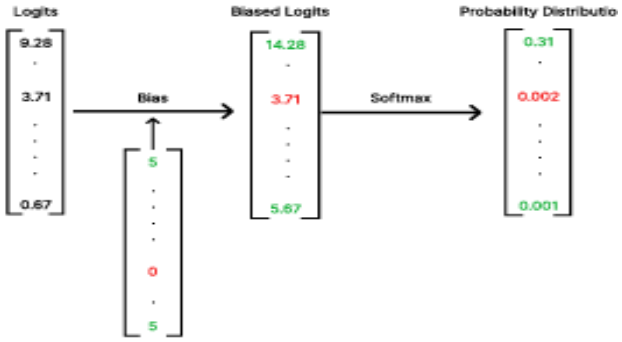


Fig. 2. Portrays the Soft Watermarking process

where the bias value is 5, in Figure 2. Visualization of adding the bias to green tokens and altering the final distribution for all the tokens.

2) Easymark

The provided watermarks are labeled as White mark, Variant mark, and Print mark. The entire family of methods is susceptible to an attack called homo glyph removal, where a homoglyph refers to a group of similar-looking characters—however, these methods of watermarking lead to no degradation in perplexity. The watermarks proposed are variations or manipulations of Unicode characters, whether it be replacing whitespace characters or replacing letters with rarer Unicode siblings.

3) Retrieval Defense

The paper proposes a retrieval-based defensive technique against paraphrasing attacks. It stores AI-generated documents in a database and determines if a document is AI-generated through a retrieval call. The method achieves 100% accuracy before paraphrasing and over 96% accuracy after strong machine paraphrasing, outperforming the Maryland watermark by 55.8%. However, the retrieval corpus matches the initial generated documents [23].

4) Other Watermarking Techniques

They propose a dynamic bias that considers the semantic meaning of the previously generated tokens, as opposed to a fixed bias like the Maryland Watermark. The technique consists of an embedding model, a custom watermark model, and a language model. The custom watermark model, T, receives embeddings as input and produces logits, PT. Paired with the logits produced by the language model, PM, I create a new distribution P ˆ M = PM +δPT, where δ weights the distribution PT [24]. A visualization of this process is provided in Figure 3:

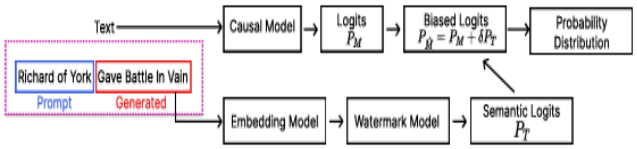


Fig. 3. Watermarking Process

C. Attacking Watermarks

Methods of removing the watermark are real threats, hence the desire for a robust watermark. Here, call attempts to remove a watermark, no matter the removal method, an attack on the watermark.

1) Word Replacement

Word replacement is a simple yet effective attack method that replaces words with synonyms instead of restructuring sentences [23]. This low-cost approach avoids intensive computation and helps remove "green tokens," reducing watermark detection likelihood [6]. This method is dependent on a Part of Speech (POS) tagger model. A POS model is trained to give grammatical structure to words. This helps us understand how to replace a given word [25]. Figure 4 provides a clear outline of the method, containing the Part of Speech tagging.

Fig. 4. Word Replacement Process Using Flair POS Tagger

2) Paraphrasing

Paraphrasing attacks target the Maryland Watermark by modifying text while preserving meaning. Without identifying 'green' tokens, paraphrasing can still reduce watermark detection. This study examines sentence-based and paragraph-based paraphrasing models, including DIPPER, which adjusts lexical and structural elements. Research shows paraphrasing significantly lowers AI-text detection accuracy, with watermarked document detection dropping from 99.8% to 30.9% after three paraphrasing iterations [26].

3) Translation-Based Paraphrasing

Translation-based watermark removal leverages existing Seq2Seq translation models to paraphrase text. As shown in Figure 5, twice-translation between languages acts as an effective paraphrasing technique, utilizing full-context decoding to enhance text transformation.

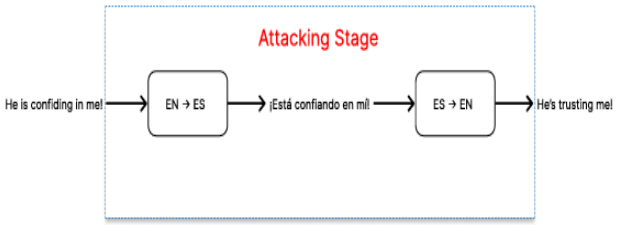


Fig. 5. Translation Attack Process (English-Spanish-English)

4) Other Attacking Techniques

Homoglyph removal and spoofing are techniques used to counter watermarking concepts. Homoglyph removal replaces characters with normal Unicode or ASCII characters, countering watermarking. Spoofing reduces credibility of watermarks, making human documents appear watermarked. Both techniques are effective but only applicable to certain watermarks, making them less discussed in this paper.

D. Approach

This paper will focus on generating documents and using three primary methods to attack them, summarizing the primary points for the rest of the paper.

- Paragraph-Based Paraphrasing: Paraphrasing the entire document, supplying the entire context [27].
- Sentence-Based Paraphrasing: Paraphrasing each sentence within a document, only supplying the context of the given sentence.
- Word Replacement: Changing words within the document to similar words.

This paper aims to understand the effectiveness of the Maryland Watermark technique against bad actors by breaking it down into four sub-questions.

E. Overarching Approach

The approach begins with creating watermarked documents through a large language model, given a dataset of prompts [28]. These watermarked documents will be paired alongside non-watermarked documents generated by the same model without the watermark. The non-watermarked documents serve as a benchmark to better understand the impact of the watermark. Consequently, these essays are passed through their chosen attacking methods. These attacking methods will be two variants of paraphrasing as well as word replacement. Figure 6 research process stages.

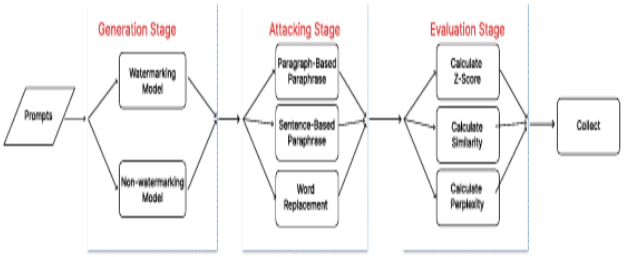


Fig. 6. Research Process Stages

F. Generation Stage

Dataset: The watermarked documents were generated using prompts from a Kaggle competition on AI-generated

text detection [29]. The DAIGT dataset, alongside human-written student essays from another Kaggle competition, was used. The dataset contains 2,421 rows with columns: id, text, and instructions, where the instructions represent student tasks, and the text column contains the corresponding essays.

Model: This paper uses Mistral-7B-Instruct-v0.2, a 7 billion parameter model, to study watermarks, a fine-tuned instruction model that has performed well in various benchmarks [30]. From Figure 7, the prompting method is designed to reflect the instruct nature required for the model.

Prompt

You are a student working on the following assignment.

Write an essay based on the following task in no more than a 100 words:
Task: Write an essay discussing the benefits of pushing yourself beyond what you have already mastered in order to grow and achieve your dreams. Use Emerson's quote as a starting point and back it up by citing real life examples from your own experiences.

Fig. 7. Prompting Method for Watermarked Essay Generation

The generation strategy used was multinomial sampling, with a maximum of 7500 new tokens, aiming to create diverse documents without using top-p sampling [31].

G. Attacking Stage

This section discusses their attacking technique for removing the Maryland Watermark using various methods, including paraphrase attack with a sentence-based and paraphrase-based paraphraser, and their word-replacement algorithm.

1) Paragraph-Based Paraphrasing

The paragraph-based paraphraser uses a Seq2Seq model fine-tuned on a 100,000-sample subset of the kPar3 dataset shows in Figure 8, which includes 6 million paraphrase pairs [32]. Paraphrasing attacks with parameters L40 and O40 are applied using top-p sampling (p=0.75). This model evaluates the effects of recursive paraphrasing on the Maryland Watermark.

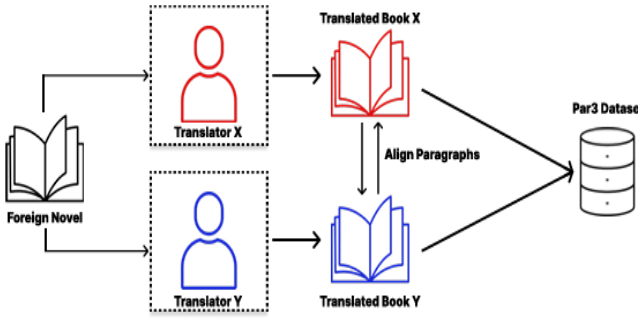


Fig. 8. Par3 Dataset Creation and Paraphrase Alignment

The kPar3 dataset contains approximately 6,000,000 paraphrase pairs, split between Google and human translations [33]. It includes documents with additional context and target portions of paraphrasing, with arguments lexical and order, to understand paraphrasing strength in the model.

2) Sentence-Based Paraphrasing

This approach paraphrases sentences independently using the ChatGPT Paraphraser, a T5-based model trained on synthetic paraphrases. Sentence tokenization is performed using NLTK, and paraphrased sentences are reassembled in order.

3) Word Replacement

The text discusses a word-replacement algorithm that uses Flair's POS Tagger to identify the grammatical nature of each word. The algorithm uses the WordNet library to recognize 155,327 distinct words and 117,597 sets of synonyms, known as Synsets. The algorithm involves tagging, selecting words based on percentage or word type, finding synonyms, choosing a synonym, and rebuilding the document. Variations are introduced, such as randomly replacing a percentage of the document with a noun-replacement approach or 25% of words replacement approach. This approach is chosen to match the green list fraction γ, allowing for effective Maryland Watermark removal. The algorithm's implementation and variations will be evaluated to determine the sufficiency of word-replacement algorithms.

H. Evolution Stage

The evaluation stage of their research involves numerically assessing a watermark's strength and associating factors. The main topics are detection, similarity, and perplexity, which measure watermark strength, paraphrase quality, and text quality, respectively [34].

1) Watermark Detection

This section outlines the Maryland detection method and key evaluation metrics used to assess watermark robustness before and after attacks. It begins with the z-score definition from [18], a statistical measure of standard deviation.

Z-score: Let $\gamma \in (0,1)$. Let T and $|s|_G$ denote size of vocabulary and number of green tokens in the document s. The Z-score are calculated as Equation (1):

$$z = \frac{|s|_G - T_\gamma}{\sqrt{T_\gamma(1-\gamma)}} \quad (1)$$

False Positive Rate: Let N and FP be the number of non-watermarked and number of false positives, respectively. I define the False Positive Rate (FPR) as follows as Equation (2):

$$FPR = \frac{FP}{N} \quad (2)$$

True Positive Rate: Let P and TP be the number of waters marked documents and the number of true positives, respectively. Then I define the True Positive Rate (TPR) as follows as Equation (3):

$$TPR = \frac{TP}{P} \quad (3)$$

2) Document Similarity

Evaluating the cost of removing the Maryland Watermark involves measuring similarity to the original document. Cosine similarity, a standard NLP metric, is applied using embeddings from a paraphrase-trained model. Since similarity scores depend on model representations, they may not fully capture paraphrasing effects. A similarity threshold of 0.76 is considered sufficient for maintaining meaning. Each attack is assessed by comparing altered documents to their originals, providing insight into meaning degradation and the overall cost of watermark removal [29].

3) Perplexity Measure

In addition to similarity measures, perplexity is used to evaluate the quality of a text and assess the cost of attacking. In NLP, perplexity gauges how likely a model is to generate a given document. Let be a document of length n and let si

denote the ith token [22]. I define the perplexity function PPL as follows as Equation (4):

$$PPL(s) = \exp\left(-\frac{1}{n}\sum_{i}^{n} log\ p(s_i|s_{<i})\right) \quad (4)$$

Where p is probability of a token given previous tokens from a given model, numerically, the model is 'unsurprised' by a generated document when the perplexity is closer to 1.

IV. Result Analysis And Discussion

The study assesses watermark detection on 1,000 documents. Recursive paraphrasing weakens detection but stabilizes. Paragraph-based paraphrasing outperforms sentence-based, reducing Type-I and Type-II errors. Word replacement fails to break the watermark. Sentence-based paraphrasing is the most effective and feasible evasion method. This analysis is completed on 500 watermarked documents alongside 500 non watermarked documents. All 1000 documents are generated through their Mistral model. Table II highlights the average document length in tokens as well as number of records for each phase of evaluation.

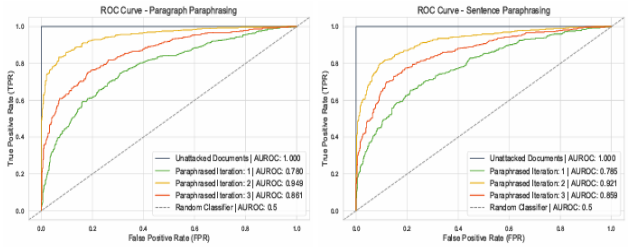
TABLE II. Document Statistics and Tokenization Details

Documents	No. Documents	No. Watermarked	No. Tokens (Mean)
Original Generated	1000	500	201.96
Paragraph-Paraphrased	996	498	154.54
Sentence-Paraphrased	996	498	191.19
Noun Word-Replaced	996	498	208.06
Percentage word-replaed	996	498	208.52

The analysis of 1000 generated documents, including 500 watermarked ones, is summarized in Table II. Various attacks, such as paraphrasing and word replacement, altered token counts, with paragraph-based paraphrasing reducing it most. Tokenization used Mistral-7B-Instruct-v0.2, with means rounded to two significant figures.

A. Recursive Paraphrasing

The research investigates the power of recursive paraphrasing in removing the Maryland Watermark. Using AUROC graphs, it is found that repeatedly paraphrasing alters detection performance. The first paraphrase is most effective, but the second iteration has a stronger watermark. This suggests that recursive paraphrasing cannot guarantee continuous degradation in watermark detection. However, the results contradict previous research, which shows iterations continually degrade detection accuracy. The paper's differences in models and prompting style remain uncertain.



(a) Paragraph-Based Paraphrasing *(b)* Sentence-Based Paraphrasing

Fig. 9. ROC Curves for Paragraph-Based and Sentence-Based Paraphrasing

The ROC curves illustrate the performance of documents after paragraph-based and sentence-based paraphrasing, highlighting AUROC values across different iterations as shown in Figure 9. An evaluation conducted on 996 documents, with half being watermarked, demonstrates lower-bound detection performance. Table III presents similarity scores across paraphrasing iterations, showing a decline in retention of the original meaning.

TABLE III. Similarity Degradation Across Paraphrasing Iterations

Documents	Similarity (↑)	
	p	***s***
First Paraphrase	0.820	0.856
Second Paraphrase	0.731	0.820
Third Paraphrase	0.671	0.785

Table III illustrates the similarity degradation across paraphrasing iterations for 996 documents, with half being watermarked. The results show a steady decline in similarity to the original text for both paragraph-based and sentence-based approaches. While the first paraphrase retains more similarity, subsequent iterations further reduce resemblance, highlighting the trade-off between effective watermark removal and text preservation. The paragraph-based paraphraser fails to maintain the original meaning in two paraphrases, indicating that repeated paraphrasing does not predictably degrade accuracy in detecting the Maryland Watermark, potentially lowering its value.

B. Sentence-Paraphrasing against Paragraph Paraphrasing

A comparison between sentence-based and paragraph-based paraphrasing reveals that the paragraph-based approach is slightly more effective in mitigating Type-I and Type-II errors. The paragraph paraphraser eliminates Type-II errors entirely while showing only a minor difference in True Positive Rate (TPR) and Type-I errors. These findings are analyzed in greater details:

TABLE IV. Performance Comparison of Paraphrase Attacking Methods

Paraphrase Attacking Method	TPR (%)	TNR (%)	Perplexity (↓)	Similarity (↑)
Paragraph-Based	1.606	100	19.530	0.820
Sentence-Based	3.815	100	24.889	0.856

The evaluation of 996 documents, with half being watermarked, assesses the effectiveness of paraphrase attacking methods, as shown in Table IV. The paragraph-based paraphraser maintains lower perplexity, while the sentence-based approach retains higher similarity. TPR and TNR are determined using a z-score threshold of 4.0, with bold values indicating the best-performing metrics. Table IV reveals that the p paraphraser provides higher-quality text, but retains higher similarity to the original document. The choice of paraphraser doesn't matter, as the p-paraphraser performs slightly better due to better model quality.

C. Word Replacement Sufficiency

Word replacement attacks are evaluated using TPR and TNR metrics at a 4.0 Kirchenbauer z-score threshold. The analysis indicates that noun replacement alone is ineffective in breaking the Maryland Watermark, and even replacing 25% of words fails to sufficiently reduce watermark detectability.

TABLE V. Evaluation of Word Replacement Methods Using TPR, TNR, and Perplexity

Replacement Method	TPR (%)	TNR (%)	Perplexity (↓)
Noun-Replacement	95.783	100	69.571
Percentage-Replacement (25%)	39.759	100	105.164

Table V displays metrics evaluated of 996 documents for each replacement method. The TPR and TNR values are complete according to the z-score of 4.0. Furthermore, the arrow denotes that it wants perplexity as a lower value, as close to 1 as possible. The percentage-replacement method is completed by replacing 25% of words in each document

The study reveals that changes in words impact z-scores, with noun-replacement and percentage-replacement attacks reducing z-scores but not sufficient. The Maryland Watermark shows no difference in z-scores within non-watermarked documents, indicating that replacement methods don't lead to spoofing attacks or undermine its credibility. Word-replacement results in a significant loss in perplexity, with human-written documents having a mean perplexity of 18.632. However, this loss in textual quality is understandable as synonyms may not always be suitable.

D. Feasibility of Attacking Techniques

The final question explores the feasibility of using attacking techniques to evade watermark detection. Word replacement is not considered feasible, while paragraph-based paraphrasing produces high-quality text and evades detection. However, this requires a 1.5B parameter causal model. The s-paraphraser is a feasible technique, as it performs similarly to paragraph-paraphraser but at a fraction of the memory costs. Sentence-based paraphrasing is a realistic technique for evading detection in academic contexts, with perplexity comparable to human-written essays.

E. Limitations

The study found that computational power was not a limitation, but the lack of appropriate metrics was a major issue. Document similarity was not an appropriate measure of paraphrase quality, and existing metrics like BLEU and ROUGE were ineffective due to their lack of alternatives. This made it difficult to provide a numerical comparison between paraphrasers for the second research question. The robustness of the Maryland Watermark was tested only on the Mistral-7B-Instruct-v0.2 model, and results may vary with other language models. Additionally, while semantic similarity and perplexity were used to assess content preservation and fluency, human evaluation was not incorporated, which could offer deeper insights into text quality and detectability. Finally, the study did not explore adversarial training or real-time evasion strategies, which may further challenge watermark resilience.

V. Conclusion And Future Work

This study demonstrates that the Maryland Watermark exhibits strong resilience against basic word-level evasion techniques, such as synonym substitution, with high true positive rates of 95.78% for noun substitution and 39.76% for 25% token replacement. However, it shows moderate vulnerability to more advanced paraphrasing strategies. Paragraph-based paraphrasing achieved the lowest perplexity (19.53), indicating higher fluency, while also causing the greatest drop in semantic similarity, thus making it the most disruptive to watermark detection. Sentence-based paraphrasing, though slightly less effective in reducing semantic fidelity, still maintained a higher perplexity (24.89) and proved to be a feasible attack vector. Recursive paraphrasing initially reduced watermark detectability but led to detection recovery in subsequent iterations, suggesting limited long-term evasion capability. Overall, the Maryland Watermark remains robust to word-level changes but requires improvement to resist paraphrasing that compromises semantic integrity. To improve the dependability and universality of attribution of AI-generated information, future research will investigate more flexible and robust watermarking methods, cross-model assessments, and linguistic robustness.

References

[1] J. Kirchenbauer, J. Geiping, Y. Wen, J. Katz, I. Miers, and T. Goldstein, “A Watermark for Large Language Models,” in *Proceedings of Machine Learning Research*, 2023.

[2] K. Y. Yoo, W. Ahn, J. Jang, and N. Kwak, “Robust Multi-bit Natural Language Watermarking through Invariant Features,” in *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 2023. doi: 10.18653/v1/2023.acl-long.117.

[3] S. Pahune and M. Chandrasekharan, “Several Categories of Large Language Models (LLMs): A Short Survey,” *Int. J. Res. Appl. Sci. Eng. Technol.*, vol. 11, no. 7, pp. 615–633, 2023, doi: 10.22214/ijraset.2023.54677.

[4] T. Munyer, A. A. Tanvir, A. Das, and X. Zhong, “DeepTextMark: A Deep Learning-Driven Text Watermarking Approach for Identifying Large Language Model Generated Text,” *IEEE Access*, 2024, doi: 10.1109/ACCESS.2024.3376693.

[5] R. Tarafdar, “Algorithms on Majority Problem,” *Univ. Missouri-Kansas City*, 2017.

[6] V. S. Sadasivan, A. Kumar, S. Balasubramanian, W. Wang, and S. Feizi, “Can AI-generated text be reliably detected?,” *arXiv Prepr. arXiv2303.11156*, 2023.

[7] S. Pandya, “Comparative Analysis of Large Language Models and Traditional Methods for Sentiment Analysis of Tweets Dataset,” *Int. J. Innov. Sci. Res. Technol.*, vol. 9, no. 12, pp. 1647–1657, 2024, doi: https://doi.org/10.5281/zenodo.14575886.

[8] N. R. Saurabh Pahune, “Healthcare: A Growing Role for Large Language Models and Generative AI,” *Int. J. Res. Appl. Sci. Eng. Technol.*, vol. 11, no. VIII, 2023, doi: 10.13140/RG.2.2.20109.72168.

[9] E. Mitchell, Y. Lee, A. Khazatsky, C. D. Manning, and C. Finn, “Detectgpt: Zero-shot machine-generated text detection using probability curvature,” in *International Conference on Machine Learning*, 2023, pp. 24950–24962.

[10] S. Murri, “Graph Database Pruning for Knowledge Representation in LLMs,” *dzone*, 2025.

[11] J. Qiang, S. Zhu, Y. Li, Y. Zhu, Y. Yuan, and X. Wu, “Natural language watermarking via paraphraser-based lexical substitution,” *Artif. Intell.*, 2023, doi: 10.1016/j.artint.2023.103859.

[12] Prity Choudhary, Rahul Choudhary and S. Garaga, “Enhancing Training by Incorporating ChatGPT in Learning Modules: An Exploration of Benefits, Challenges, and Best Practices,” *Int. J. Innov. Sci. Res. Technol.*, vol. 9, no. 11, 2024.

[13] J. Qiu, W. Han, X. Zhao, and S. Long, “Evaluating Durability : Benchmark Insights into Multimodal Watermarking,” *J. Data-centric Mach. Learn. Res.*, 2024.

[14] K. S. Saurabh Pahune, Zahid Akhtar, Venkatesh Mandapati, “The Importance of AI Data Governance in Large Language Models,” *Preprints*, 2025.

[15] S. Rastogi and D. Pruthi, “Revisiting the Robustness of Watermarking to Paraphrasing Attacks,” Nov. 2024, doi: arXiv:2411.05277.

[16] Z. Zhang *et al.*, “Stealing Watermarks of Large Language Models via Mixed Integer Programming,” in *2024 Annual Computer Security Applications Conference (ACSAC)*, IEEE, Dec. 2024, pp. 46–60. doi: 10.1109/ACSAC63791.2024.00021.

[17] N. R. Barman *et al.*, “The Brittleness of AI-Generated Image Watermarking Techniques: Examining Their Robustness Against Visual Paraphrasing Attacks,” *arXiv*, Aug. 2024, doi: arXiv:2408.10446.

[18] H. Chang, H. Hassani, and R. Shokri, “Watermark Smoothing Attacks against Language Models,” *Under Rev. as a Conf. Pap. ICLR*, Jul. 2024, doi: 10.48550/arXiv.2407.14206.

[19] B. Y. Idrissi, M. Millunzi, A. Sorrenti, L. Baraldi, and D. Dementieva, “Temperature Matters : Enhancing Watermark Robustness Against Paraphrasing Attacks,” 2024.

[20] R. Sennrich, B. Haddow, and A. Birch, “Neural machine translation of rare words with subword units,” *arXiv Prepr. arXiv1508.07909*, 2015.

[21] R. Sato, Y. Takezawa, H. Bao, K. Niwa, and M. Yamada, “Embarrassingly simple text watermarks,” *arXiv Prepr. arXiv2310.08920*, 2023.

[22] J. Kirchenbauer, J. Geiping, Y. Wen, J. Katz, I. Miers, and T. Goldstein, “A watermark for large language models,” in *International Conference on Machine Learning*, 2023, pp. 17061–17084.

[23] K. Krishna, Y. Song, M. Karpinska, J. Wieting, and M. Iyyer, “Paraphrasing evades detectors of ai-generated text, but retrieval is an effective defense,” *Adv. Neural Inf. Process. Syst.*, vol. 36, pp. 27469–27500, 2023.

[24] A. Liu *et al.*, “A survey of text watermarking in the era of large language models,” *ACM Comput. Surv.*, vol. 57, no. 2, pp. 1–36, 2024.

[25] Z. He *et al.*, “Can watermarks survive translation? on the cross-lingual consistency of text watermark for large language models,” *arXiv Prepr. arXiv2402.14007*, 2024.

[26] A. Akbik, D. Blythe, and R. Vollgraf, “Contextual string embeddings for sequence labeling,” in *Proceedings of the 27th International Conference on Computational Linguistics*, 2018, pp. 1638–1649.

[27] A. Q. Jiang *et al.*, “Mixtral of experts,” *arXiv Prepr. arXiv2401.04088*, 2024.

[28] C. de L. Pataca and P. D. P. Costa, “Speech-modulated typography,” in *IUI ’20: Proceedings of the 25th International Conference on Intelligent User Interfaces*, OSF, pp. 139–143. doi: 10.1145/3377325.337752.

[29] J. Wieting, K. Gimpel, G. Neubig, and T. Berg-Kirkpatrick, “Paraphrastic representations at scale,” *arXiv Prepr. arXiv2104.15114*, 2021.

[30] A. B. Hou, J. Zhang, Y. Wang, D. Khashabi, and T. He, “k-semstamp: A clustering-based semantic watermark for detection of machine-generated text,” *arXiv Prepr. arXiv2402.11399*, 2024.

[31] S. Atawneh and H. Aljehani, “Phishing Email Detection Model Using Deep Learning,” *Electron.*, 2023, doi: 10.3390/electronics12204261.

[32] K. Thai *et al.*, “Exploring document-level literary machine translation with parallel paragraphs from world literature,” *arXiv Prepr. arXiv2210.14250*, 2022.

[33] Y. Tay *et al.*, “Scale efficiently: Insights from pre-training and fine-tuning transformers,” *arXiv Prepr. arXiv2109.10686*, 2021.

[34] L. R. Kirkendall and T. H. Atkinson, “What we do and don’t know about New World pinhole borers (Coleoptera, Curculionidae, Platypodinae),” *Nor. J. Entomol.*, 2024.