An Analysis into the Robustness of Watermarking Generative Text

Sam Jackson

2520998J

Introduction – Misinformation, ChatGPT and Hallucination

NEWSLETTERS · EYE ON AI

OpenAI's Sora has left AI experts either enthused or skeptical. It's left most everyone else terrified

BY SAGE LAZZARO
February 22, 2024 at 5:44 PM GMT

A smartphone displaying a screen shot from one of the demo videos OpenAI made public to show off the capabilities of its new text-to-video generating AI, Sora. The new AI model thrilled some. But it terrified many others.
CFOTO/FUTURE PUBLISHING VIA GETTY IMAGES

Hello and welcome to Eye on AI.

It's been a week since OpenAI unveiled Sora, its new text-to-video generative AI model it says can turn short text prompts into strikingly realistic videos up to a minute long. The videos shared thus far have been received as thoroughly impressive and a giant leap for AI video generation (said with some reservations to account for the fact that OpenAI hasn't demonstrated the model actually working or released a technical report). But today, I'm diving into a different through line of the reaction to Sora, and that's fear.

Related Video

Forbes

FORBES > LEADERSHIP > LEADERSHIP STRATEGY

ChatGPT: Concerns, Fears And Opportunities

Shep Hyken Contributor
CX & Customer Service Expert, Researcher, Speaker and Author

Follow

Jul 30, 2023, 08:00am EDT

← All Open Letters

Pause Giant AI Experiments: An Open Letter

We call on all AI labs to immediately pause for at least 6 months the training of AI systems more powerful than GPT-4.

Signatures
33708

Add your signature

Published
March 22, 2023

[1]: ChatGPT: Concerns, Fears and Opportunities - Accessed 20/03/24
[2]: OpenAI's Sora is terrifying people – Accessed 20/03/24
[3]: Pause Giant AI Experiments – Accessed 20/03/24

"As a large language model..."

Famously, ChatGPT begins some responses with "As a large language model...". This portion of text acts as a degree of authentication and certification that it was written by ChatGPT.

Obviously, this text can be removed with a few backspaces. But what if it was made harder to remove? This is how we introduce *watermarking*.

Watermarking is a way to imprint a certifying factor into content.

With the introduction of watermarks, we want to know one thing:

<u>What does it take to remove a watermark?</u>

[1]: 100 Dollar Bill - Accessed 20/03/24

Background

Language Generation

Watermarking Documents

Attacking Documents

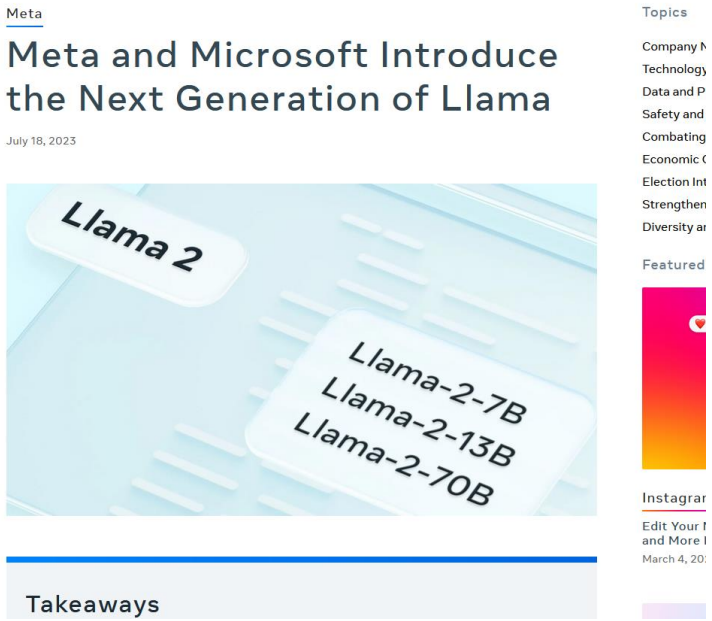
Causal Language Generation

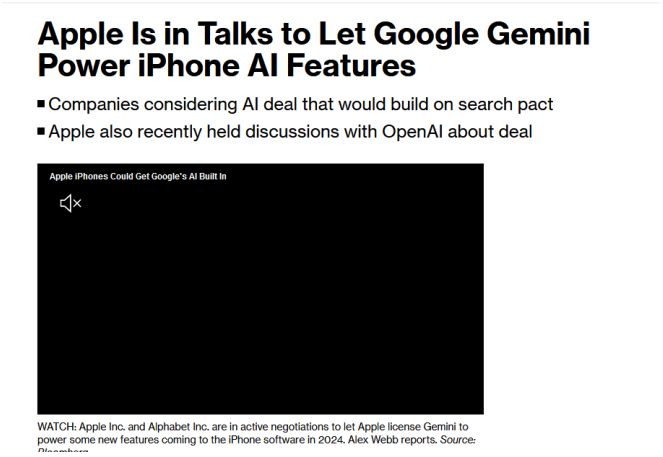
We discuss language generation with regards to **Causal** Language Generation.

In generated text, each token is produced one at a time, given the prior tokens.

A language model spits out a probability distribution, from which the generated token is chosen.

Famous models of causal language generation include Gemini, ChatGPT, Mistral and Llama.





[1]: Mistral & Microsoft – Accessed 21/03/24
[2]: Facebook - Llama-2 – Accessed 21/03/24
[3]: Apple & Gemini – Accessed 21/03/24

Watermarks from Maryland and Beyond

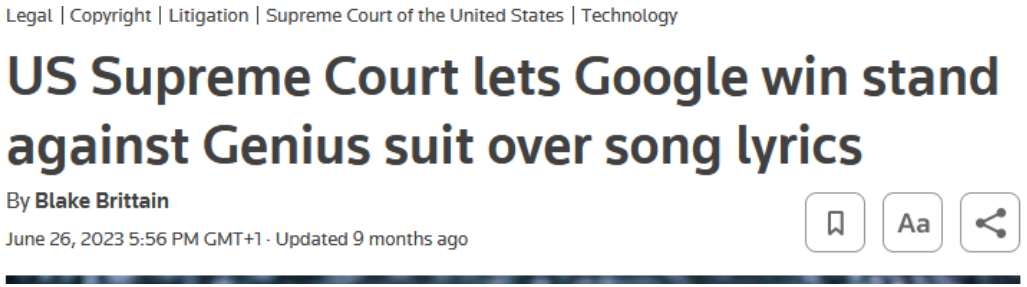
On our journey to understand watermarks, we choose to focus on the progenitor of watermarking techniques.

A text watermark is designed to achieve three primary criteria:

- **Robust** – Capable of withstanding attempts to remove the watermark.
- **Agnostic** – Information beyond generated text is not required for detection.
- **Impercetible** – The watermark should not be visible, be it a change in quality or a written signature.

The Maryland Watermark [1], by Kirchenbauer, proposes dynamically biasing the probability distribution towards certain tokens. As certain tokens appear, we can detect the use of the Large Language Model

Other simpler techniques include hiding watermarks amongst special Unicode characters. Google was caught out with this technique, by Genius, in 2019 [2]. Sato proposes a family of such watermarks in his paper on simple watermarks [3].


Legal | Copyright | Litigation | Supreme Court of the United States | Technology

US Supreme Court lets Google win stand against Genius suit over song lyrics

By **Blake Brittain**

June 26, 2023 5:56 PM GMT+1 · Updated 9 months ago

[1]: Maryland Watermark – Accessed 20/03/24
[2]: Google Stealing Genius Lyrics – Accessed 20/03/24
[3]: Embarrassingly Simple Watermarks – Accessed 20/03/24

Do watermarks exist forever?

What if I wanted to remove the Maryland Watermark? Is this possible?

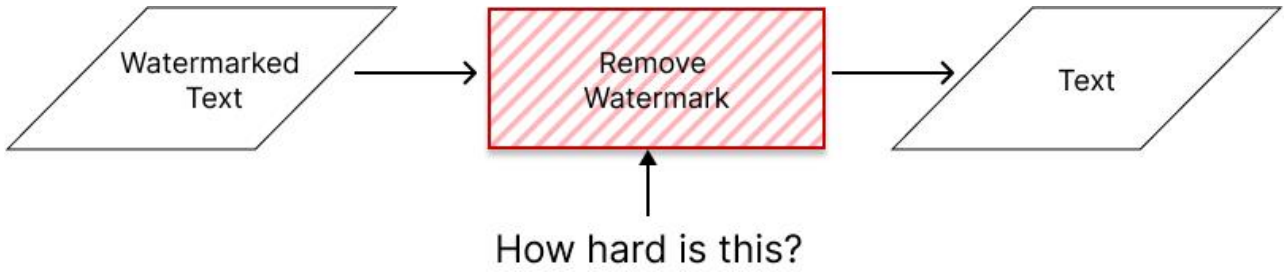
The simplest method of removal is by removing all of the generated text. We are interesting in being able to remove the watermark with efficiency.

An efficient watermark removal would retain the following properties of the original document:

- **Clarity** – The original meaning is not lost.
- **Quality** – Removing the watermark does not degrade the quality of the generated text.
- **Cost** – The price to remove the watermark is not unreasonable.

Researched methods of removal include:

- Paraphrase – Rewrite the text in your own words whilst maintaining the meaning.
- Word Replacement – Replace words, without changing the structure.



Paragraph vs Sentence - Paraphasing

Paragraph-Based Paraphrasing & Sentence-Based Paraphrasing

Using paragraphs as paraphrasing has the potential to incorporate further meaning.

Sentence

I could see a elephant with a telescope

Paragraph

The telescope, clutched in my hands, helped me see animals. I could see a elephant with a telescope

What do we want to find out?

In our paper, we are questioning the **Robust** property of the Maryland Watermark.

We frame our search as a single problem: "To what extent is the Maryland Watermark robust against bad actors?"

This question is broken into 4 research-questions:

- **RQ1**: Does paraphrasing recursively degrade accuracy in detection of the Maryland Watermark?
- **RQ2**: Is sentence-based paraphrasing more effective than paragraph-based paraphrasing when dealing with removal of the Maryland Watermark?
- **RQ3**: Is a low-cost, word-replacement algorithm sufficient to remove the Maryland Watermark?
- **RQ4**: Are the attacking methods feasible for use within an academic context with respect to the Maryland Watermark?

In this presentation, we will only discuss results for **RQ3**.

Implementation

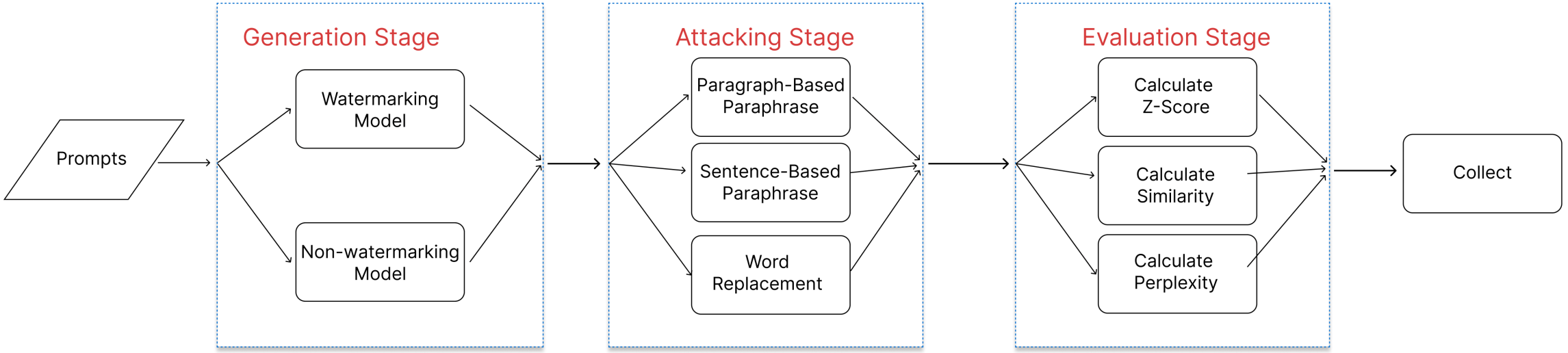
Maryland Watermark

Attacking Methods

Evaluation Metrics

PROCESS

Displayed below is our research approach, designed to help us answer our questions



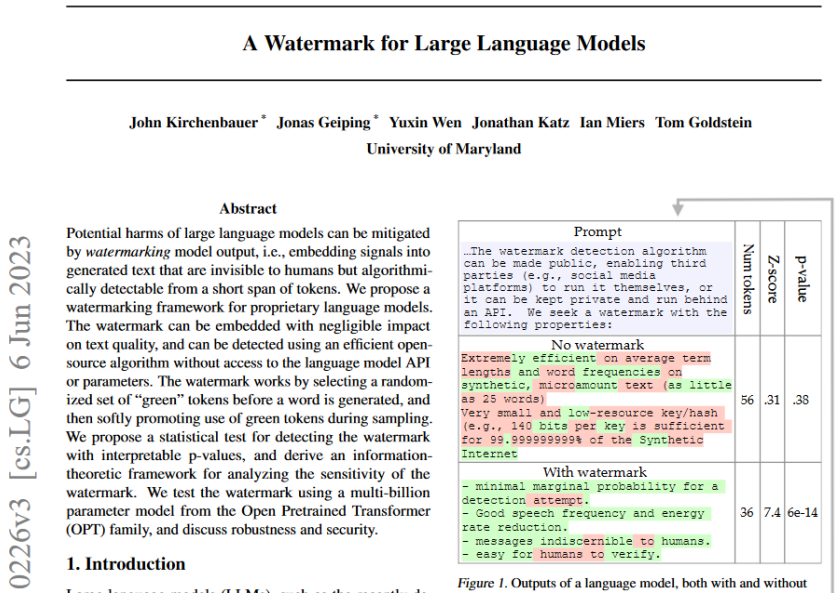
We refer to attempts to remove the watermark as *attacking* the watermark.

Making our watermark

Given our research questions, we implement the Maryland Watermark.

We choose a moderate strength watermark, using parameters recommended by Kirchenbauer [1].

Our generative text model is *Mistral-7B-Instruct-v0.2*, a successful model on many benchmarks [2].



A Watermark for Large Language Models

John Kirchenbauer * Jonas Geiping * Yuxin Wen Jonathan Katz Ian Miers Tom Goldstein
University of Maryland

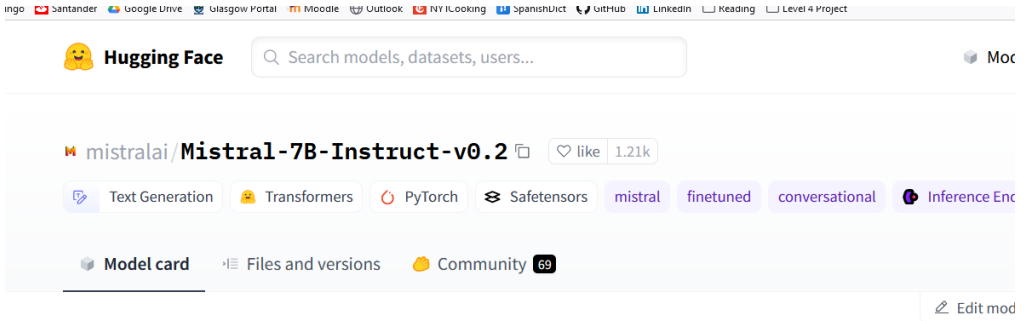
Abstract

Potential harms of large language models can be mitigated by *watermarking* model output, i.e., embedding signals into generated text that are invisible to humans but algorithmically detectable from a short span of tokens. We propose a watermarking framework for proprietary language models. The watermark can be embedded with negligible impact on text quality, and can be detected using an efficient open-source algorithm without access to the language model API or parameters. The watermark works by selecting a randomized set of "green" tokens before a word is generated, and then softly promoting use of green tokens during sampling. We propose a statistical test for detecting the watermark with interpretable p-values, and derive an information-theoretic framework for analyzing the sensitivity of the watermark. We test the watermark using a multi-billion parameter model from the Open Pretrained Transformer (OPT) family, and discuss robustness and security.

1. Introduction

Prompt	Num tokens	Z-score	p-value
...The watermark detection algorithm can be made public, enabling third parties (e.g., social media platforms) to run it themselves, or it can be kept private and run behind an API. We seek a watermark with the following properties:			
No watermark			
Extremely efficient on average term lengths and word frequencies on synthetic, microamount text (as little as 25 words) Very small and low-resource key/hash (e.g., 140 bits per key is sufficient for 99.9999999999% of the Synthetic Internet	56	.31	.38
With watermark			
- minimal marginal probability for a detection attempt. - Good speech frequency and energy rate reduction. - messages indiscernible to humans. - easy for humans to verify.	36	7.4	6e-14

Figure 1. Outputs of a language model, both with and without



[1]: Maryland Watermark – Accessed 20/03/24
[2]: Mistral Model – Accessed 20/03/24

Attacking

Paragraph-Paraphrasing:

Created a smaller model and finetuned based off a model described by Krishna in their retrieval paper [1].

Sentence-Paraphrasing:

Noting the lack of paraphrasers amongst current literature, I use a popular one amongst HuggingFace. ChatGPT paraphraser [2].

Word-Replacement:

Use a sequence tagger [3], providing grammatical structure, and then use WordNet [4] to provide synonyms and replace certain words.

Both Noun-replacements and percentage replacement.

[1]: Krishna | Retrieval Paper – Accessed 21/03/24
[2]: ChatGPT Paraphraser – Accessed 21/03/24
[3]: Flair POS Tagger – Accessed 21/03/24
[4]: WordNet – Accessed 21/03/24

Evaluation – From One Metric to Another

Z-Score:

Z-Score is a statistical test where, in our case, it is used to determine if a text is watermarked.

TPR:

The ratio of watermarked documents correctly detected as watermarked.

$$TPR = \frac{TP}{Actual\ Positive} = \frac{TP}{TP + FN}$$

FPR:

The ratio of human documents incorrectly detected as watermarked

$$FNR = \frac{FN}{Actual\ Positive} = \frac{FN}{TP + FN}$$

$$TNR = \frac{TN}{Actual\ Negative} = \frac{TN}{TN + FP}$$

$$FPR = \frac{FP}{Actual\ Negative} = \frac{FP}{TN + FP}$$

Similarity:

Using an embedding model, we can calculate similarity between documents.

Similarity is in the range [0,1] and helps determine whether the original meaning of a document has been lost.

Perplexity:

The textual quality of a model, calculated with a neutral language model. Perplexity measures a degree of 'surprise' that a model produced text.

Analysis

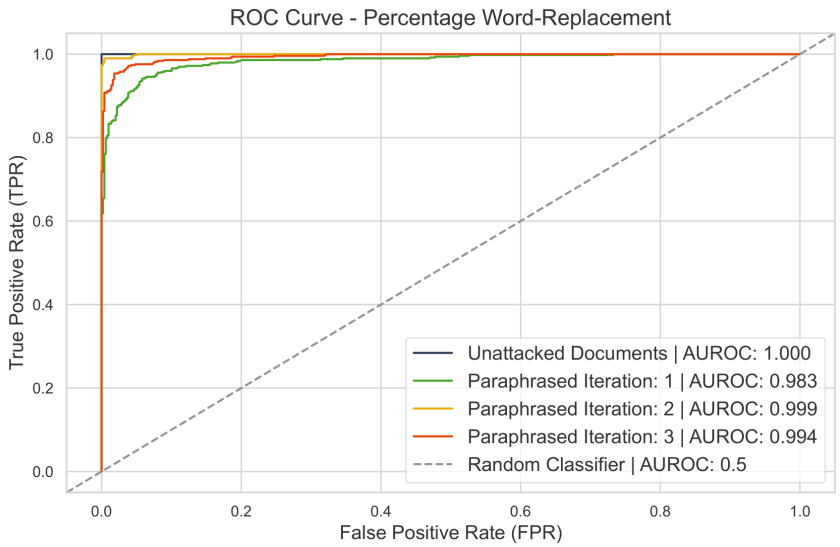
RQ1: Does paraphrasing recursively degrade accuracy in detection of the Maryland Watermark?

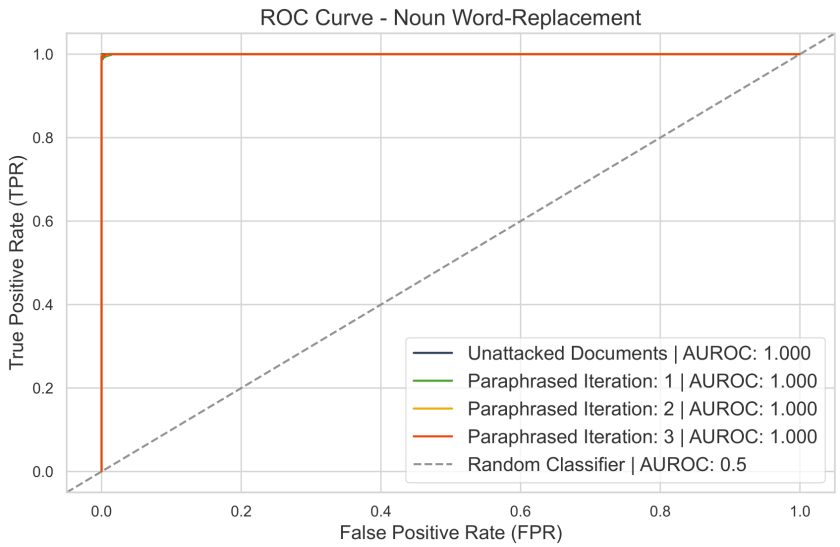
RQ2: Is sentence-based paraphrasing more effective than paragraph-based paraphrasing when dealing with removal of the Maryland Watermark?

RQ3: Is a low-cost, word-replacement algorithm sufficient to remove the Maryland Watermark?

RQ4: Are the attacking methods feasible for use within an academic context with respect to the Maryland Watermark?

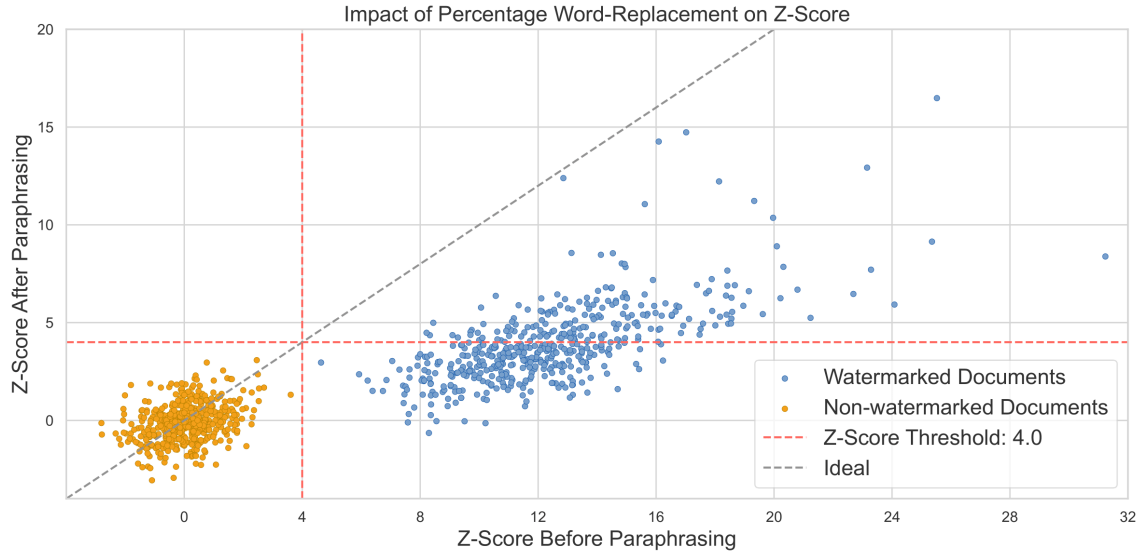
Results – RQ3 - AUROC

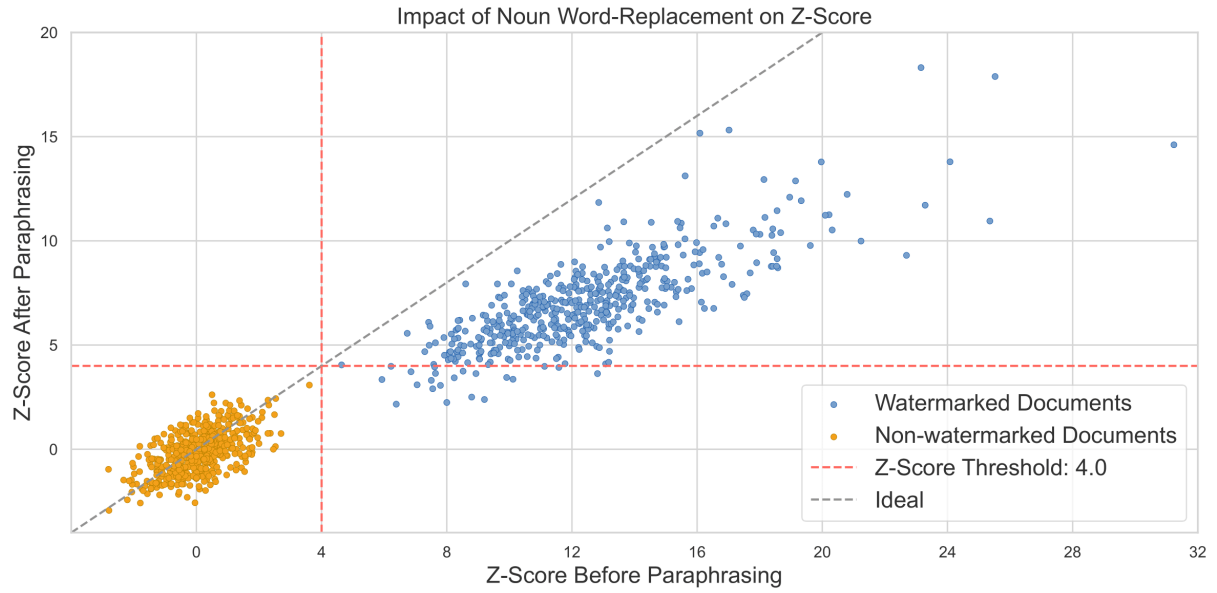




Replacement Method	TPR (%)	TNR (%)	Perplexity (↓)
Noun-Replacement	95.783	100	69.571
Percentage-Replacement (25%)	39.759	100	105.164

Results – RQ3


Impact of Percentage Word-Replacement on Z-Score
Z-Score After Paraphrasing
Z-Score Before Paraphrasing
Watermarked Documents
Non-watermarked Documents
Z-Score Threshold: 4.0
Ideal


Impact of Noun Word-Replacement on Z-Score
Z-Score After Paraphrasing
Z-Score Before Paraphrasing
Watermarked Documents
Non-watermarked Documents
Z-Score Threshold: 4.0
Ideal

Conclusion

Questions Answered

Limitations

Future Work

Conclusion

Summary:

We have answered each of our research questions succinctly.

- Recursively paraphasing is not reliable.
- The use of sentence-based or paragraph-based paraphrasing is not important.
- Word Replacement is insufficient as a removal technique.
- Sentence-based paraphrasing is feasible technique for academic use

Future Work:

We propose two further research questions.

RQ5: Does the use of a sliding-window for z-score calculations provide greater accuracy for detection of documents that contain AI-generated text?

RQ6: Do the previously mentioned paraphrasing attacks succeed against a semantic-influenced watermarking technique?

Limitations:

We discuss the computational limitations as well the metric limitations.