Lost in Overlap: Exploring Logit-based Watermark Collision in LLMs

Yiyang Luo[*]
Nanyang Technological University
lawrenceluoyy@outlook.com

Ke Lin[*]
Tsinghua University
leonard.keilin@gmail.com

Chao Gu[*]
Univ. of Science and Technology of China
guch8017@mail.ustc.edu.cn

Jiahui Hou
Univ. of Science and Technology of China
jhhou@ustc.edu.com

Lijie Wen
Tsinghua University
wenlj@tsinghua.edu.cn

Ping Luo
Tsinghua University
luop@tsinghua.edu.cn

Abstract

The proliferation of large language models (LLMs) in generating content raises concerns about text copyright. Watermarking methods, particularly logit-based approaches, embed imperceptible identifiers into text to address these challenges. However, the widespread usage of watermarking across diverse LLMs has led to an inevitable issue known as watermark collision during common tasks, such as paraphrasing or translation. In this paper, we introduce watermark collision as a novel and general philosophy for watermark attacks, aimed at enhancing attack performance on top of any other attacking methods. We also provide a comprehensive demonstration that watermark collision poses a threat to all logit-based watermark algorithms, impacting not only specific attack scenarios but also downstream applications.

1 Introduction

As the quality of text produced by large language models (LLMs) advances, it addresses numerous practical challenges while raising many new issues. In particular, the widespread generation of text by LLMs on the Internet may increase the spread of rumors and raise concerns about text copyright (Megías et al., 2021; Tang et al., 2023). Consequently, the identification and classification of machine-generated text have become critically significant. Watermarking techniques for LLMs can help to tackle these problems, leading to their rising importance in ongoing conversations and attracting increasing interest globally.

Text watermarking involves embedding distinctive, imperceptible identifiers (watermarks) into written content. Nowadays, most methods are logit-based (Kirchenbauer et al., 2023a; Liu et al., 2023a; Zhao et al., 2023; Kuditipudi et al., 2023; Hu et al., 2023): they manipulate the output logits of LLMs during the generation process using distinct but consistently logit-based strategies to embed watermarks successfully. Utilizing the power of LLMs ensures that such adjustings for probabilistic distribution are seamlessly integrated into the generated content without compromising the overall quality or coherence of the text (Lin et al., 2024). These watermark methods are sophisticatedly designed to be robust yet discreet, ensuring content integrity and ownership preservation without compromising readability or meaning.

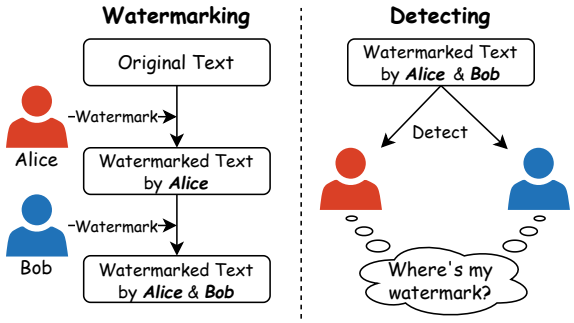


Figure 1: Illustration of watermark collisions.

However, as watermarking techniques proliferate, watermark collision becomes inevitable with the increasing application of watermarks. The term **watermark collision** can be defined as instances where the text contains multiple watermarks simultaneously (Fig. 1). This is particularly inevitable during tasks that may require the collaboration of multiple LLMs, such as paraphrasing and translation. While many methods (Liu et al., 2023a; Zhao et al., 2023; Kuditipudi et al., 2023) claim resilience against paraphrase attacks, none have been specifically tested for watermark collisions. Hence, we examine the underlying mathematical principles, employ watermark collision as an attack strategy for all watermarking techniques that utilize logits, and evaluate its effectiveness in conjunction

[*]Equal Contribution.
[†]Code and data are available at https://github.com/AInnovateLab/watermark-collision.

Findings of the Association for Computational Linguistics:
NAACL 2025, pages 620–637

with multiple traditional attack methods enhanced through the incorporation of watermark collision.

Our Contributions. In summary, this paper proposes **a new watermark attack philosophy for all logit-based watermarks** in LLMs. Our contributions are as follows:

- We propose a novel philosophy for watermark attacks that can effectively remove existing watermarks from text. This approach can be integrated with various traditional attack methods to enhance their performance.
- We find that the strength of overlapping watermarks impacts detection performance. Upstream and downstream watermarks generally compete for detection accuracy, with one being stronger and the other weaker.
- We discuss the vulnerability of watermarking techniques caused by watermark collisions.

2 Related Work

Text watermarking. Modern text watermarking techniques can be classified into two categories: modification-based and generation-based watermarks (Liu et al., 2024). Modification-based watermarking, also known as watermarks for existing text, consists of altering an existing text to create a watermarked text. Most of the modification-based techniques can be classified as either lexical (Topkara et al., 2006b; Yang et al., 2022; Munyer and Zhong, 2023; Sato et al., 2023) or syntactic methods (Atallah et al., 2001; Topkara et al., 2006a; Meral et al., 2009), based on rules, classical machine learning or deep neural models.

LLM watermarking. While modification-based techniques (Abdelnabi and Fritz, 2020; Yang et al., 2022) modify the text and preserve its semantics, generation-based methods apply watermarks into the text generation process to achieve better results, enabling smoother integration with LLMs. Watermark injection can be carried out during either the training phase or the inference phase.

During the training of LLMs, watermarks are inserted into the training data to intentionally alter the results of LLMs for certain inputs (Liu et al., 2023b; Sun et al., 2022). The main objective of training time watermarking is to protect dataset copyrights from unauthorized usage (Tang et al., 2023; Sun et al., 2023). Despite being able to embed watermarks in LLMs, training-time watermarking has significant limitations, including limited payload capacity, restricted trigger conditions, and significant training overhead.

For inference-time watermarking, Kirchenbauer et al. (2023a) proposed a logit-based greenlist mechanism based on prior token hashes. Liu et al. (2023a) introduced a watermark model to generate semantic-preserving logits during text generation. Zhao et al. (2023) simplified the Kirchenbauer et al.'s scheme by using a fixed Green-Red split and achieved greater robustness. Christ et al. (2023); Kuditipudi et al. (2023); Fu et al. (2024) aim to design watermark techniques that are more robust and secure. Yoo et al. (2023); Boroujeny et al. (2024) introduce watermarking techniques with increased payload capacity for arbitrary binary data. Zhu et al. (2024) enhances the efficiency and quality of watermarking by embedding dual secret patterns.

Even though these methods have been designed to be more robust against attacks such as paraphrase attacks (Kirchenbauer et al., 2023b), back-translation attacks (He et al., 2024) and mask-and-fill attacks (Lyu et al., 2023), these attacks often use unwatermarked LLMs, e.g. DIPPER (Krishna et al., 2023) and GPT-3.5 (Brown et al., 2020). Prior research in the pre-LLM era has mentioned potential risks associated with multiple watermarks (Tanha et al., 2012). Nevertheless, the effects of one watermarking technique on another in the context of LLM watermarking remain unclear, which is the motivation for this study.

3 Method

3.1 Principle of Watermark Collision

The detection process for logit-based watermarking methods relies on the null hypothesis testing. A well-known example is the null hypothesis of KGW (Kirchenbauer et al., 2023a):

$$H_0 : \textit{The text sequence is generated with no knowledge of the red list rule.} \quad (1)$$

Since the words of red list are chosen randomly, a natural writer is expected to sample words both from red and green list, whereas the watermarked model produces words only from green list. In practice, effective detection typically requires the text to contain sufficient words from a runtime-generated green list. This requirement ensures that

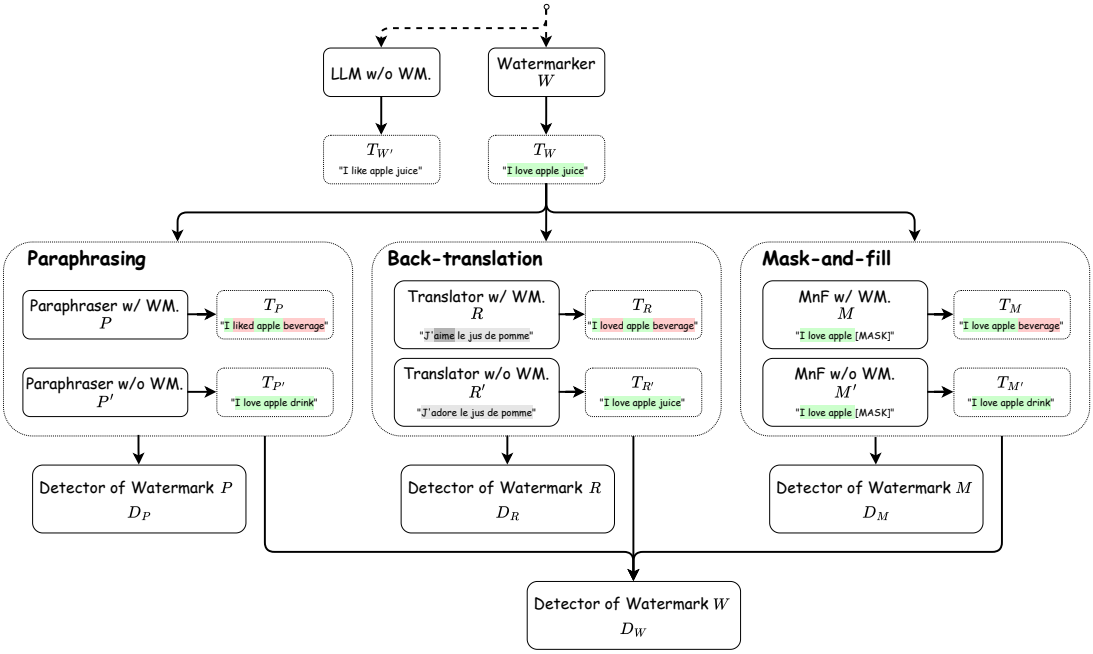


Figure 2: The collision pipeline. T_W denotes text with the first watermark W, where T_C denotes text with dual watermarks from a different collider $C \in \{P, R, M\}$. Unwatermarked text generated from W and C is denoted as $T_{W'}$ and $T_{C'}$. T_C and $T_{C'}$ are then examined by D_W and D_C to determine the presence of watermark W and C. Texts in red and green are visualization samples of the red-green list showing the original watermark W.

the generated text adheres to a specific probability distribution. The detection algorithm checks whether the distribution of words in the text conforms to the green-list distribution.

Let $\mathbf{P}(T)$ represent the probability distribution of a certain text T. The detection process verifies if $\mathbf{P}(T)$ follows the null hypothesis H_0, which can be formulated as the follows, where $\boldsymbol{\delta}(\cdot)$ is the dirichlet function:

$$\boldsymbol{\delta}[T \text{ is watermarked}] = \begin{cases} 0 & \mathbf{P}(T) \approx \mathbf{P}(H_0) \\ 1 & \mathbf{P}(T) \not\approx \mathbf{P}(H_0) \end{cases} \tag{2}$$

From a probabilistic perspective, a watermark w can be detected only when text T follows a watermarked distribution $\mathbf{P}_w$. However, each watermarking method has an independent null hypothesis H_0 and therefore creates different word distributions in the generation process. When an additional watermark w' is applied, it imposes a new distribution $\mathbf{P}_{w'}$ on the text. If more watermarks are added, a series of distributions are obtained: $(\mathbf{P}_{w^{(0)}}, \mathbf{P}_{w^{(1)}}, \mathbf{P}_{w^{(2)}}, \ldots, \mathbf{P}_{w^{(n)}})$, where $\mathbf{P}_{w^{(0)}} = \mathbf{P}_w$. Introducing each new watermark modifies the word probabilities in T, thus altering the overall distribution. Specifically, any distinct watermarks $w^{(i)}$ and $w^{(j)}$ should have different distributions $\mathbf{P}_{w^{(i)}}$ and $\mathbf{P}_{w^{(j)}}$ to ensure that they are not incorrectly detected by each other. As a result, the text T no longer strictly follows the original distribution $\mathbf{P}(T)$, nor does it fully conform to any of the subsequent distributions $\mathbf{P}_{w^{(0)}}, \mathbf{P}_{w^{(1)}}, \mathbf{P}_{w^{(2)}}, \ldots, \mathbf{P}_{w^{(n)}}$. Therefore, the combination of multiple watermarks can be described as a transformation of the original watermark into an entangled one:

$$\mathbf{P}_{\text{entangled}} = f(\mathbf{P}_{w^{(0)}}, \mathbf{P}_{w^{(1)}}, \mathbf{P}_{w^{(2)}}, \ldots, \mathbf{P}_{w^{(n)}}) \tag{3}$$

Here, f represents the complex transformation function resulting from the sequential application of multiple watermarks. This new distribution $\mathbf{P}_{\text{entangled}}$ is not merely a simple combination but a new complex distribution that emerges from the (indirect) interaction of all applied watermarks.

Since the detection process relies on identifying patterns consistent with $\mathbf{P}_{w^{(i)}}$, the introduction of $\mathbf{P}_{\text{entangled}}$ causes the final text distribution to no longer conform to any of these patterns. The detection algorithm may not be able to detect the watermark in the text due to a shift in word distribution from $P_{w^{(i)}}$, leading to *watermark collisions*, whereas standard attacks are random, unstable, and less effective.

3.2 Pipeline Design

To prove the existence of watermark collisions, we design pipelines with three main components: *wa-*

termarker, *colliders*, and *detectors*.

3.2.1 Watermarker

Watermarker W generates watermarked texts T_W by using a language model (LM) to create content based on a specific corpus as context. As illustrated in Fig. 2, we first produce the watermarked text data T_W with *Watermarker* W. Additionally, we generate unwatermarked text $T_{W'}$ using the same context and prompt as T_W for further comparisons. Section 4 and Appendix A provide details regarding the watermarker setup.

3.2.2 Colliders

Colliders C are designed to attack the watermark created by the *watermarker* using collision techniques. There are three distinct *colliders* that apply such collision attacks through traditional attack methods, namely paraphraser, back-translator, and mask-and-filler.

Paraphrase Collider. *Paraphraser* P rephrases the watermarked texts T_W with different watermarks, i.e. generated by different methods or keys, to generate paraphrased text data T_P, which are intended to contain dual watermarks simultaneously. Furthermore, we also generate texts T'_P using the same paraphraser but without a watermark, denoted as P', for further comparison.

Back-translation Collider. *Translator* R translates the watermarked texts T_W to other languages and then translates back to their original language with watermarks. As shown in Fig. 2, we first paraphrase the original text data T_W using *Watermarker* W. Then the text data will be translated and back-translated by *Translator* R with different watermark settings to generate text data T_R which are intended to contain dual watermarks simultaneously. Furthermore, we also generate texts T'_R using the same translator but without a watermark, denoted as R', for further comparison.

It is important to note that the KGW-based method is essentially ineffective against back-translation attacks due to the inability to capture the contextual semantics. Thus, they have been excluded from this pipeline. We choose French as the pivot language in the back-translation. Details are in Section 4.

Mask-and-fill Colliders. *Mask-and-filler* (MnF) M is specifically designed for mask-and-fill attacks. The MnF attack method is commonly used with masked language models, e.g., BERT-based models. For our study, we opted for RoBERTa$_{\texttt{LARGE}}$ as the base model. As shown in Fig. 2, we first generate watermarked text data T_W using *Watermarker* W. Then the text data will be mask-and-filled by *MnF* M with different watermark settings to generate text data T_M which are intended to contain dual watermarks simultaneously. Additionally, we create texts T'_M by applying the same MnF but excluding the watermark, denoted as M'. These will be used as baseline texts for comparison.

3.2.3 Detectors

As demonstrated in Fig.2, four detectors are tailored to identify a specific type of watermark. Detector D_P targets watermarks in paraphrasers, D_R focuses on those in translators, and D_M is for watermarks in the MnF process. Detector D_W aims to identify the original watermark embedded by the watermarker. By comparing the results from these detectors, we can assess the effectiveness of the attacks with or without additional watermarks.

4 Experiments

4.1 Experiment Setup

Settings. We utilize the C4 dataset (Raffel et al., 2020) as the context for text generation with a maximum of 128 tokens. For watermarker and paraphraser, we employ LLaMA-2-13B (Touvron et al., 2023), Qwen2-7B (Bai et al., 2023) and OPT-1.3B (Zhang et al., 2022). For back-translator, we use LLaMA-2-13B and Qwen2-7B. For mask-and-filler, we exclusively utilize RoBERTa$_{\texttt{LARGE}}$ (Liu et al., 2019) as the masked language model. We only present the results of LLaMA-2-13B here and refer to the Appendix B for more details.

Watermarks. Our experiments were conducted using several previous watermark methods as baselines. The watermark strength of each method is tuned separately for `weak` and `strong` watermarks. Methods involved include: (a) **KGW** from Kirchenbauer et al. (2023a), $d = 2, 5$ for weak and strong settings; (b) **PRW** from Zhao et al. (2023), $s = 2, 5$ for weak and strong settings. (c) **SIR** from Liu et al. (2023a), $d = 2, 5$ for weak and strong settings; Note that we use subscript $\cdot_{\texttt{wk}}$ and $\cdot_{\texttt{sg}}$ to refer to weak and strong watermarks, respectively.

Some methods are not selected for specific reasons. **UBW** from Hu et al. (2023) asserting that their approach lacks resilience against paraphrase

W \ P	∅	P'	KGW_{wk}		PRW_{wk}		SIR_{wk}	
	D_W	D_W	D_W	D_P	D_W	D_P	D_W	D_P
KGW_{wk}	99.90	71.65	52.80	19.80	41.10	48.00	3.40	90.09
PRW_{wk}	95.40	49.25	37.00	28.20	26.60	41.20	22.30	79.56
SIR_{wk}	87.90	59.05	55.74	25.10	41.05	19.80	/	/

(a) weak W, weak P

W \ P	∅	P'	KGW_{wk}		PRW_{wk}		SIR_{wk}	
	D_W	D_W	D_W	D_P	D_W	D_P	D_W	D_P
KGW_{wk}	99.90	71.65	4.10	97.40	6.20	99.90	0.20	92.77
PRW_{wk}	95.40	49.25	14.60	96.70	9.30	99.90	29.40	91.91
SIR_{wk}	87.90	59.05	12.45	96.70	13.22	97.50	/	/

(b) weak W, strong P

W \ P	∅	P'	KGW_{sg}		PRW_{sg}		SIR_{sg}	
	D_W	D_W	D_W	D_P	D_W	D_P	D_W	D_P
KGW_{sg}	100.00	81.95	68.20	21.50	56.00	44.40	9.00	79.26
PRW_{sg}	99.70	75.05	67.90	18.30	44.20	41.70	32.00	65.31
SIR_{sg}	94.30	67.60	61.55	21.00	45.58	14.00	/	/

(c) strong W, weak P

W \ P	∅	P'	KGW_{sg}		PRW_{sg}		SIR_{sg}	
	D_W	D_W	D_W	D_P	D_W	D_P	D_W	D_P
KGW_{sg}	100.00	81.95	7.00	94.10	14.80	99.70	0.80	90.52
PRW_{sg}	99.70	75.05	14.60	96.70	11.20	99.30	37.10	79.00
SIR_{sg}	94.30	67.60	10.98	97.40	12.58	97.10	/	/

(d) strong W, strong P

Table 1: TPR of the *paraphrased* text T_P with dual watermarks when FPR $= 1\%$. W and P represent the watermarker and paraphraser, respectively. D_W and D_P represent the detector of the watermarker and paraphraser. ∅ indicates that no paraphrasing process is applied to the text, and its corresponding column represents the result of using D_W to detect watermark W in T_W. P' represents paraphrasing T_W without watermark, as mentioned in Fig. 2.

attacks, thereby rendering it unsuitable for this experiment. **RDW** from Kuditipudi et al. (2023) may be ineffective when subjected to a key different from their recommended configuration, thereby failing to meet the experimental requirements that necessitate distinct key settings.

4.2 Evaluation Metrics

Given the rapid development of the watermarking field, there is currently no consensus on metrics, with different methods employing varied evaluation criteria (Tu et al., 2023). Moreover, some approaches utilize a detection threshold to categorize text as watermarked, while others differ. Implementing all possible metrics for each method is impractical and biased.

Following the approach of Zhao et al. (2023), we opt for a fair evaluation by avoiding the influence of threshold settings. Our final detection metrics include false positive rates (**FPR**) and true positive rates (**TPR**). We specifically set FPR values at 1%, 5%, and 10%, adjusting the detector's thresholds accordingly. This ensures a consistent and unbiased assessment of watermarking methods. We only present results when FPR $= 1\%$ here and refer to Appendix B for comprehensive results.

4.3 Experimental Results & Analysis

Watermark collision is compatible with the majority of existing attacks. Tables 1, 2, and 5 demonstrate that watermark collision is feasible for all selected attacks, including paraphrase, back-translation, and mask-and-fill attacks. Watermark collision is commonly found in attacks involving auto-regressive text generation methods such as paraphrasing and back-translation. Mask-and-fill attacks are ineffective as they cannot completely change the distribution of words in a sentence.

Watermark collision will not degrade the text quality. As shown in Tab. 3 and 4, we present perplexity before and after attacks, both with and without collisions. We use LLaMA-2-13B as the backbone for perplexity calculation, which is the same model for the initial generated text. As evidence, text quality remains largely stable post-attack, and most collisions did not result in significant declines in text quality, indicating the potential value of collision as an attack methodology. Detailed semantic analysis examples of the text quality can be found in Appendix B.

Watermark attacks with watermarks collision tend to be stronger than those without. In Table 1 and 2, we present the detection accuracy for various baseline watermark algorithms with and without the occurrence of watermark collision (T_P and T'_P, respectively). A noteworthy decline in detection accuracy is observed when watermarks are introduced in the context of traditional attacks such as paraphrase attacks and back-translation attacks. According to the settings in Table 1 and 2, there is a strong competition between overlapping watermarks. As one watermarker attempts to maintain its detection accuracy, the others' detection accuracy decreases. SIR-SIR is not listed since the SIR does not allow the user to choose a key to create a

W \ R	$\varnothing$	R'	KGW_{wk}		PRW_{wk}		SIR_{wk}	
	D_W	D_W	D_W	D_R	D_W	D_R	D_W	D_R
KGW_{wk}	99.90	44.80	41.90	9.20	34.20	13.70	31.40	5.15
PRW_{wk}	95.40	32.90	34.10	7.50	21.80	12.70	25.20	5.73
SIR_{wk}	87.90	5.60	4.75	7.70	6.12	11.50	/	/

(a) weak W, weak R

W \ R	$\varnothing$	R'	KGW_{wk}		PRW_{wk}		SIR_{wk}	
	D_W	D_W	D_W	D_R	D_W	D_R	D_W	D_R
KGW_{wk}	99.90	44.80	28.50	69.50	18.50	58.50	3.90	93.50
PRW_{wk}	95.40	32.90	23.00	68.00	14.30	54.60	10.80	92.87
SIR_{wk}	87.90	5.60	4.28	67.40	4.40	41.60	/	/

(b) weak W, strong R

W \ R	$\varnothing$	R'	KGW_{wk}		PRW_{wk}		SIR_{wk}	
	D_W	D_W	D_W	D_R	D_W	D_R	D_W	D_R
KGW_{sg}	100.00	69.30	67.80	8.10	61.90	13.60	55.30	4.09
PRW_{sg}	99.70	63.90	64.80	9.60	55.20	10.20	50.60	3.56
SIR_{sg}	94.30	3.00	1.74	9.40	3.82	6.70	/	/

(c) strong W, weak R

W \ R	$\varnothing$	R'	KGW_{wk}		PRW_{wk}		SIR_{wk}	
	D_W	D_W	D_W	D_R	D_W	D_R	D_W	D_R
KGW_{sg}	100.00	69.30	50.00	68.90	38.80	53.00	7.50	91.70
PRW_{sg}	99.70	63.90	50.10	67.90	36.60	42.40	20.10	90.44
SIR_{sg}	94.30	3.00	4.15	74.40	2.51	22.50	/	/

(d) strong W, strong R

Table 2: TPR of the *back-translated* text T_R with dual watermarks when FPR $= 1\%$. Similar to Table 1. Data annotated with underline indicate abnormal data points.

different distribution, and therefore the watermark collision will not occur on SIR-SIR. However, SIR watermarks are still vulnerable to collision attacks using other watermark methods.

It should also be noted that different watermarking methods behave differently during competition. KGW appears to be less competent than the other two methods. SIR, however, shows significant collisions even in weak paraphraser settings (Column SIR of Tab. 1a & 1c), while PRW exhibits extreme collisions in strong paraphraser settings (Column PRW of Tab. 1b & 1d).

However, some anomalies can be observed in back-translation (Table 2), where the TPR increases after a collision. These anomalies occur because certain watermarks may have similar distributions in specific contexts, rendering them ineffective and making attacks on them pointless: When two watermark methods exhibit similar distributions, the likelihood of collision decreases, leading to an increase in the True Positive Rate (TPR). However, in fact, this is a degraded performance for a certain watermark method: It is a challenge to identify the origin of the watermark. In practical scenarios, if the origin of the watermark cannot be determined, it essentially means we cannot ascertain which entity embeds the watermark and makes the watermark meaningless. For example, if two KGW-based watermark methods utilize the *a similar but not the same* red-green list, paraphrasing words into other similar words may accidentally reinforce the watermark within the text, thereby enhancing the TPR for both watermarks and lead to confusion on which entity apply watermark on the text. Consequently, such cases are considered **failures** of watermark methods and do not concern us.

MnF exhibits a similar trend as the watermark collision intensifies, however, it experiences less impact compared to other colliders. We note that attacks on MnF are less effective because the unmasked words retain most of the context and keep the watermark unaltered. Therefore, the TPR in these cases remains nearly the same.

Nonetheless, it is observed that when considering the z-score of H_0 as stated in (Kirchenbauer et al., 2023a), the influence of watermark collision persists. In Table 5, we present the z-scores for each corresponding method. Degradation of the z-score is observed after collision attacks compared to those without, indicating the effectiveness of collisions. Besides, the mask rate has a greater impact on the detection process. If the mask rate is low, MnF is less likely to have a significant effect, thus explaining why these attacks have minimal impact on MnF. Furthermore, the SIR method is excluded from MnF experiments, as it prioritizes semantic factors and is less susceptible to MnF attacks.

4.4 Multi-round Collision

To further enhance the performance of these attacks, multi-round collisions can be applied. We tested the performance of multi-round attacks both with and without watermark collisions, and the results are presented in Fig. 3. In each experiment, we use the same paraphraser to assess the chain effect caused by watermark collisions. For various watermark methods, the TPR of each watermark detection decreases after multi-round collision attacks. As explained in the previous section, applying multi-round collisions causes the word distribu-

W \ P	$\varnothing$	P'	KGW_{wk}	PRW_{wk}	SIR_{wk}
KGW_{wk}	7.36	6.96	8.60	6.58	12.28
PRW_{wk}	6.92	6.69	9.77	6.24	12.15
SIR_{wk}	10.14	7.15	11.04	7.09	/

(a) weak W, weak P

W \ P	$\varnothing$	P'	KGW_{sg}	PRW_{sg}	SIR_{sg}
KGW_{wk}	7.36	6.96	15.67	5.05	14.08
PRW_{wk}	6.92	6.69	13.85	5.22	10.25
SIR_{wk}	10.14	7.15	13.31	5.12	/

(b) weak W, strong P

W \ P	$\varnothing$	P'	KGW_{wk}	PRW_{wk}	SIR_{wk}
KGW_{sg}	13.18	7.44	14.16	8.39	13.55
PRW_{sg}	8.31	7.61	10.53	6.54	12.00
SIR_{sg}	12.30	7.72	12.75	7.80	/

(c) strong W, weak P

W \ P	$\varnothing$	P'	KGW_{sg}	PRW_{sg}	SIR_{sg}
KGW_{sg}	13.18	7.44	12.80	5.60	12.03
PRW_{sg}	8.31	7.61	11.15	4.91	9.22
SIR_{sg}	12.30	7.72	11.29	4.99	/

(d) strong W, strong P

Table 3: PPL of the *paraphrased* T_P with dual WMs.

W \ P	$\varnothing$	P'	KGW_{wk}	PRW_{wk}	SIR_{wk}
KGW_{wk}	7.36	8.82	9.00	8.29	10.16
PRW_{wk}	6.92	8.21	8.52	8.02	9.39
SIR_{wk}	10.14	9.82	10.37	9.14	/

(a) weak W, weak R

W \ P	$\varnothing$	P'	KGW_{sg}	PRW_{sg}	SIR_{sg}
KGW_{wk}	7.36	8.82	11.86	9.63	33.35
PRW_{wk}	6.92	8.21	11.15	8.84	30.57
SIR_{wk}	10.14	9.82	13.39	9.92	/

(b) weak W, strong R

W \ P	$\varnothing$	P'	KGW_{wk}	PRW_{wk}	SIR_{wk}
KGW_{sg}	13.18	12.74	12.88	11.41	14.91
PRW_{sg}	8.31	9.02	9.29	8.33	11.11
SIR_{sg}	12.30	10.05	10.91	8.72	/

(c) strong W, weak R

W \ P	$\varnothing$	P'	KGW_{sg}	PRW_{sg}	SIR_{sg}
KGW_{sg}	13.18	12.74	17.58	11.83	36.84
PRW_{sg}	8.31	9.02	12.59	9.60	27.98
SIR_{sg}	12.30	10.05	14.78	9.18	/

(d) strong W, strong R

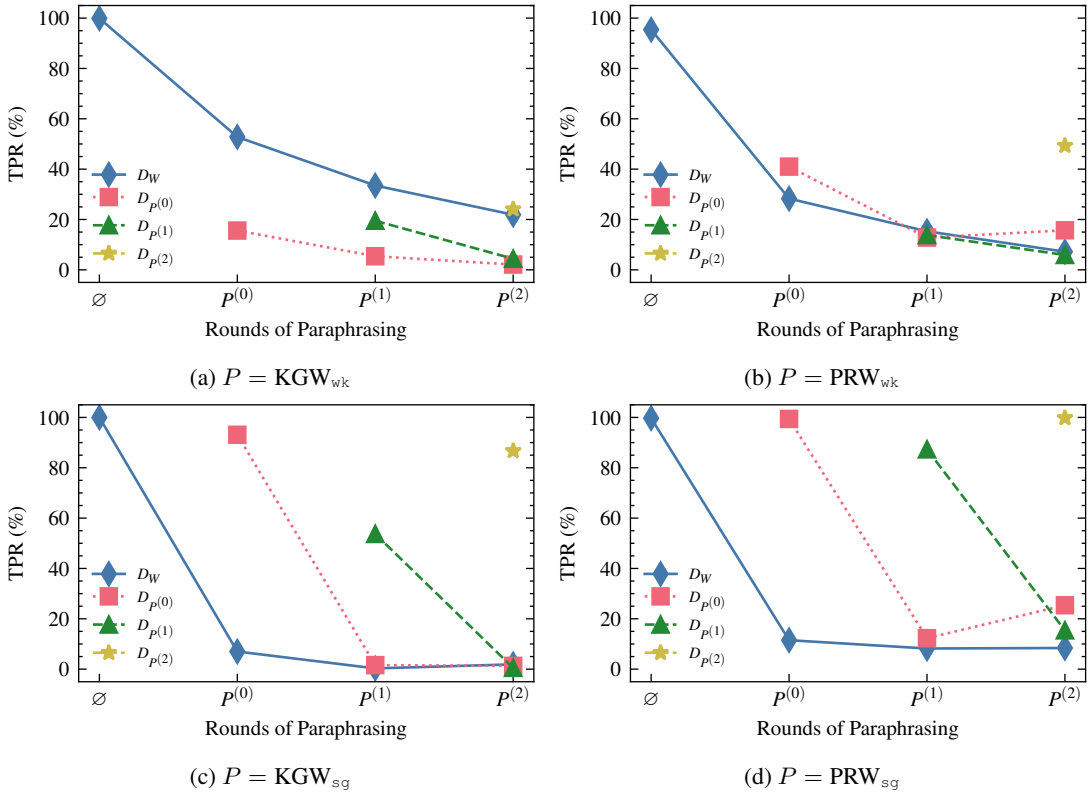
Table 4: PPL of the *back-translated* T_R with dual WMs.

tion to deviate more significantly from subsequent distributions. The stronger the watermark, the less likely it is for the multi-round watermark to coexist with others.

5 Possible Application

Malicious attacks based on watermark collisions. Previous works (Kirchenbauer et al., 2023b,a; Kuditipudi et al., 2023) have introduced several attacks, such as copy-paste attacks and paraphrase attacks, but most have shown their robustness and security against at least some of these attacks. Our study, however, provides a feasible method of constructing effective attacks using watermark collisions. For example, text with a KGW_{wk} watermark can be detected with a 44.8% TPR after a paraphrase attack without a watermark. However, if the paraphrase attack is conducted with another KGW_{wk}, the detection TPR drops to 41.9%. If the attack is done with a KGW_{sg}, the detection TPR further decreases to 28.5%, as presented in Table 1a and 1b. The use of colliders with strong watermarks could easily erase existing watermarks, resulting in greater vulnerability to watermarking.

Detection of existing watermarks using collisions between watermarks of different strengths. In Table 1 and 2, we demonstrate that weak watermarks can still be easily applied to unwatermarked text. It is, however, much more difficult to apply a weak watermark to text that has already been watermarked (Tab. 1 & 2). For example, KGW can be applied to plaintext with a success rate of 99.90%, but can be applied to a SIR-watermarked text with a probability not exceeding 25.10%, as presented in Table 1a and 1c. This provides a simple probabilistic method of detecting watermarks without the need to know their details. When adding a weak watermark to a sentence is difficult, it is more likely to have an existing one in the original sentence.



(a) $P = \text{KGW}_{\text{wk}}$ (b) $P = \text{PRW}_{\text{wk}}$

(c) $P = \text{KGW}_{\text{sg}}$ (d) $P = \text{PRW}_{\text{sg}}$

Figure 3: Multi-round TPR of *paraphrased* text under a series of paraphrase attacks by the same type of paraphraser with different watermarks. ∅ represents the original detection TPR before paraphrasing. A sequence of paraphrasers $(P^{(0)}, P^{(1)}, P^{(2)}, \dots)$ is applied consecutively to the generated text from the preceding paraphraser.

W \ M	∅	KGW_{wk}	PRW_{wk}	KGW_{sg}	PRW_{sg}
KGW_{wk}	$4.14_{\pm 1.57}$	$3.88_{\pm 1.51}$	$3.62_{\pm 1.52}$	$3.52_{\pm 1.51}$	$3.42_{\pm 1.56}$
PRW_{wk}	$5.14_{\pm 1.51}$	$4.89_{\pm 1.52}$	$4.64_{\pm 1.59}$	$4.35_{\pm 1.60}$	$4.44_{\pm 1.57}$
KGW_{sg}	$6.72_{\pm 1.95}$	$6.37_{\pm 1.95}$	$6.09_{\pm 1.99}$	$5.97_{\pm 1.87}$	$5.86_{\pm 1.95}$
PRW_{sg}	$7.82_{\pm 1.70}$	$7.62_{\pm 1.74}$	$7.35_{\pm 1.81}$	$7.08_{\pm 1.87}$	$7.07_{\pm 1.82}$

Table 5: Z-scores of the *mask-and-filled* text T_M with dual watermarks with a mask rate of 0.6 for masked language modeling tasks. ∅ means no MnF is applied, i.e., the original detection results of z-scores.

6 Discussion

The LLM watermarking technique is currently undergoing rapid development, with many foundational aspects not yet implemented for practical use regularly. However, our research demonstrates that when watermarks collide, it can significantly hinder the performance of the watermark when applied in real-world situations. A list of predictable risks in practical applications is provided:

- **API Tracing**: LLM providers can use watermarking techniques on their LLM API to prevent unauthorized use. However, if the watermarked output of LLMs is sent to other providers with watermarks for further processing, the upstream watermarks will not be effective in tracing the use of the upstream APIs.
- **Black-box Detection**: As we discussed in the Experiments Section, the watermark collisions could perform black-box detection of any existing watermarks in any text. Users would experience distrust when they become aware of the presence of watermarks. Hackers may attempt to bypass the watermarks.

7 Conclusion

In this study, we examine how overlapping watermarks in the same text can decrease the accuracy of both upstream and downstream watermark detection. We propose the use of **watermark collisions as an attacking philosophy** and therefore emphasize that watermark collisions may compromise the validity and security of all logit-based watermarks. We conduct experiments by integrating various watermarkers and colliders, assessing the text quality before and after collisions, as well as with and without collisions, to demonstrate that collision can

enhance common attacks and text quality remains consistent when compared to attacks without collisions. We hope our work will increase awareness of potential threats to LLM watermarking.

Limitations

Our approach indeed demonstrates the potential collision between existing watermark techniques. Nevertheless, we conduct our experiments only on paraphrasing, back-translation, and mask-and-fill tasks to simulate watermark collisions. Broader but less related tasks, such as question answering, have not yet been tested.

Furthermore, the range of models chosen is limited. The models chosen include LLaMA-2-13B, Qwen2-7B, OPT-1.3B, and RoBERTa$_{\texttt{LARGE}}$ as the colliders, while certain other models are tailored for particular tasks. Experiments on more open-sourced models could further enhance the conclusion of our paper.

References

Sahar Abdelnabi and Mario Fritz. 2020. Adversarial watermarking transformer: Towards tracing text provenance with data hiding. *2021 IEEE Symposium on Security and Privacy (SP)*, pages 121–140.

Mikhail J Atallah, Victor Raskin, Michael Crogan, Christian Hempelmann, Florian Kerschbaum, Dina Mohamed, and Sanket Naik. 2001. Natural language watermarking: Design, analysis, and a proof-of-concept implementation. In *Information Hiding: 4th International Workshop, IH 2001 Pittsburgh, PA, USA, April 25–27, 2001 Proceedings 4*, pages 185–200. Springer.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Massieh Kordi Boroujeny, Ya Jiang, Kai Zeng, and Brian Mark. 2024. Multi-bit distortion-free watermarking for large language models. *arXiv preprint arXiv:2402.16578*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Miranda Christ, Sam Gunn, and Or Zamir. 2023. Undetectable watermarks for language models. *arXiv preprint arXiv:2306.09194*.

Jiayi Fu, Xuandong Zhao, Ruihan Yang, Yuansen Zhang, Jiangjie Chen, and Yanghua Xiao. 2024. Gumbelsoft: Diversified language model watermarking via the gumbelmax-trick. *arXiv preprint arXiv:2402.12948*.

Zhiwei He, Binglin Zhou, Hongkun Hao, Aiwei Liu, Xing Wang, Zhaopeng Tu, Zhuosheng Zhang, and Rui Wang. 2024. Can watermarks survive translation? on the cross-lingual consistency of text watermark for large language models. *arXiv preprint arXiv:2402.14007*.

Zhengmian Hu, Lichang Chen, Xidong Wu, Yihan Wu, Hongyang Zhang, and Heng Huang. 2023. Unbiased watermark for large language models. *arXiv preprint arXiv:2310.10669*.

John Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, and Tom Goldstein. 2023a. A watermark for large language models. In *Proc. 40th Int. Conf. Mach. Learn.*, volume 202 of *Proc. Mach. Learn. Res.*, pages 17061–17084. PMLR.

John Kirchenbauer, Jonas Geiping, Yuxin Wen, Manli Shu, Khalid Saifullah, Kezhi Kong, Kasun Fernando, Aniruddha Saha, Micah Goldblum, and Tom Goldstein. 2023b. On the reliability of watermarks for large language models. *arXiv preprint arXiv:2306.04634*.

Kalpesh Krishna, Yixiao Song, Marzena Karpinska, John Wieting, and Mohit Iyyer. 2023. Paraphrasing evades detectors of ai-generated text, but retrieval is an effective defense. *arXiv preprint arXiv:2303.13408*.

Rohith Kuditipudi, John Thickstun, Tatsunori Hashimoto, and Percy Liang. 2023. Robust distortion-free watermarks for language models. *arXiv preprint arXiv:2307.15593*.

Ke Lin, Yiyang Luo, Zijian Zhang, and Luo Ping. 2024. Zero-shot generative linguistic steganography. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5168–5182, Mexico City, Mexico. Association for Computational Linguistics.

Aiwei Liu, Leyi Pan, Xuming Hu, Shiao Meng, and Lijie Wen. 2023a. A semantic invariant robust watermark for large language models. *arXiv preprint arXiv:2310.06356*.

Aiwei Liu, Leyi Pan, Yijian Lu, Jingjing Li, Xuming Hu, Xi Zhang, Lijie Wen, Irwin King, Hui Xiong, and Philip S. Yu. 2024. A survey of text watermarking in the era of large language models. *arXiv preprint arXiv:2312.07913*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Yixin Liu, Hongsheng Hu, Xuyun Zhang, and Lichao Sun. 2023b. Watermarking text data on large language models for dataset copyright protection. *arXiv preprint arXiv:2305.13257*.

Mingzhi Lyu, Yi Huang, and Adams Wai-Kin Kong. 2023. Adversarial attack for robust watermark protection against inpainting-based and blind watermark removers. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 8396–8405.

David Megías, Minoru Kuribayashi, Andrea Rosales, and Wojciech Mazurczyk. 2021. Dissimilar: Towards fake news detection using information hiding, signal processing and machine learning. In *Proceedings of the 16th International Conference on Availability, Reliability and Security*, pages 1–9.

Hasan Mesut Meral, Bülent Sankur, A Sumru Özsoy, Tunga Güngör, and Emre Sevinç. 2009. Natural language watermarking via morphosyntactic alterations. *Computer Speech & Language*, 23(1):107–125.

Travis Munyer and Xin Zhong. 2023. Deeptextmark: Deep learning based text watermarking for detection of large language model generated text. *arXiv preprint arXiv:2305.05773*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Ryoma Sato, Yuki Takezawa, Han Bao, Kenta Niwa, and Makoto Yamada. 2023. Embarrassingly simple text watermarks. *arXiv preprint arXiv:2310.08920*.

Zhensu Sun, Xiaoning Du, Fu Song, and Li Li. 2023. Codemark: Imperceptible watermarking for code datasets against neural code completion models. In *Proceedings of the 31st ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, pages 1561–1572.

Zhensu Sun, Xiaoning Du, Fu Song, Mingze Ni, and Li Li. 2022. Coprotector: Protect open-source code against unauthorized training usage with data poisoning. In *Proceedings of the ACM Web Conference 2022*, pages 652–660.

Ruixiang Tang, Qizhang Feng, Ninghao Liu, Fan Yang, and Xia Hu. 2023. Did you train on my dataset? towards public dataset protection with cleanlabel backdoor watermarking. *ACM SIGKDD Explorations Newsletter*, 25(1):43–53.

Maryam Tanha, Seyed Dawood Sajjadi Torshizi, Mohd Taufik Abdullah, and Fazirulhisyam Hashim. 2012. An overview of attacks against digital watermarking and their respective countermeasures. In *Proceedings Title: 2012 International Conference on Cyber Security, Cyber Warfare and Digital Forensic (CyberSec)*, pages 265–270.

Mercan Topkara, Umut Topkara, and Mikhail J Atallah. 2006a. Words are not enough: sentence level natural language watermarking. In *Proceedings of the 4th ACM international workshop on Contents protection and security*, pages 37–46.

Umut Topkara, Mercan Topkara, and Mikhail J Atallah. 2006b. The hiding virtues of ambiguity: quantifiably resilient watermarking of natural language text through synonym substitutions. In *Proceedings of the 8th workshop on Multimedia and security*, pages 164–174.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Shangqing Tu, Yuliang Sun, Yushi Bai, Jifan Yu, Lei Hou, and Juanzi Li. 2023. Waterbench: Towards holistic evaluation of watermarks for large language models. *arXiv preprint arXiv:2311.07138*.

Xi Yang, Jie Zhang, Kejiang Chen, Weiming Zhang, Zehua Ma, Feng Wang, and Nenghai Yu. 2022. Tracing text provenance via context-aware lexical substitution. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11613–11621.

KiYoon Yoo, Wonhyuk Ahn, and Nojun Kwak. 2023. Advancing beyond identification: Multi-bit watermark for language models. *arXiv preprint arXiv:2308.00221*.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. 2022. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.

Xuandong Zhao, Prabhanjan Ananth, Lei Li, and Yu-Xiang Wang. 2023. Provable robust watermarking for ai-generated text. *arXiv preprint arXiv:2306.17439*.

Chaoyi Zhu, Jeroen Galjaard, Pin-Yu Chen, and Lydia Y Chen. 2024. Duwak: Dual watermarks in large language models. *arXiv preprint arXiv:2403.13000*.

Appendix

A Pipeline Setup

Datasets. To generate the watermarked text T_W, the C4 dataset is used as the context. A total of 1000 watermarked textual samples are generated from the selected context. To ensure that only text with an apparent watermark is selected for T_W, a z-score threshold is set during the generation. For KGW and PRW, the z-threshold is set to 4.0. For SIR, the z-threshold is set to 0.0.

Hyperparameters. We specifically designate 2024 as the key for the watermarker (if applicable) and 2023 as the key for the paraphraser (if applicable) to vary the key for watermark collision within the same watermark algorithm (e.g. **KGW-KGW**). The specific hyperparameters of each watermarking method are as follows:

- **KGW**: For weak settings, we set the green list size $\gamma = 0.25$, hardness parameter $\delta = 2.0$ and the seeding scheme is `selfhash`. For strong settings, we set the green list size $\gamma = 0.25$, hardness parameter $\delta = 5.0$ and the seeding scheme is `selfhash`.

- **SIR**: We employ the `context` mode. For weak settings, We set chunk_length $= 10$, $\gamma = 0.5$, watermark strenght $\delta = 2.0$. For strong settings, we set chunk_length $= 10$, $\gamma = 0.5$, watermark strength $\delta = 5.0$.

- **PRW**: For weak settings, we set the green list size $\gamma = 0.25$, watermark strength $\delta = 2.0$. For strong settings, we set the green list size $\gamma = 0.25$, watermark strength $\delta = 5.0$.

```
«SYS»
Assume you are a helpful assistant.
Your job is to paraphrase the given
text.
«/SYS»

[INST]{INPUT_TEXT}[/INST]

You're welcome!  Here's a
paraphrased version of the original
message:
```

Figure 4: The paraphrase prompt template for LLaMA-2 paraphraser.

Prompts. We formulate a prompt tailored for LLaMA-2-13B, enabling it to proficiently paraphrase the given content. The prompt template is shown in Fig. 4, Fig. 5, and Fig. 7.

B Experimental Results

Tables 6 and 9 show the TPR results of detection when utilizing LLaMA-2-13B as the base model of watermarker. Table 7 shows the TPR results of detection when utilizing OPT-1.3B as the base model of watermarker. Table 10 shows the TPR results of detection when utilizing OPT-1.3B as the base model of watermarker.

Tables 11 and 12 also present several examples of the watermarked texts under different settings.

Collision can be observed across different base models. This observation is supported by the use of LLaMA-2-13B, Qwen2-7B, and OPT-1.3B as the base models, as illustrated in Fig. 6. The findings suggest that watermark collision is inevitable across different base models, proving its universal applicability as a methodology.

The semantics of paraphrasing is mostly maintained in weak settings, while strong colliders preserve it to some degree. Table 8 shows the similarity of sentence embeddings across various settings, measured by cosine similarity using the `all-MiniLM-L6-v2` model from the `sentence-transformer` library.

C Scientific Artifacts

The licenses for all the watermarking methods are listed below: KGW (Apache 2.0 Licence), SIR (MIT Licence), PRW (MIT Licence). The licenses for models are listed below: LLaMA-2-13B (LLAMA 2 Community License), Qwen2-7B-Instruct (Apache 2.0 Licence), OPT-1.3B (OPT LICENSE).

```
<|im_start|>system
You are a helpful assistant.  Your
job is to paraphrase the given
text.<|im_end|>
<|im_start|>user
{INPUT_TEXT}<|im_end|>
<|im_start|>assistant
You're welcome!  Here's a
paraphrased version of the original
message:
```

Figure 5: The paraphrase prompt template for Qwen2 paraphraser.

Watermarker	KGW$_{weak}$				KGW$_{strong}$			
Paraphraser	KGW$_{weak}$		KGW$_{strong}$		KGW$_{weak}$		KGW$_{strong}$	
FPR	D_W	D_P	D_W	D_P	D_W	D_P	D_W	D_P
1%	52.80	19.80	4.10	97.40	68.20	21.50	7.00	94.10
5%	66.60	34.20	14.30	99.60	72.80	32.70	16.20	97.90
10%	71.00	41.50	22.10	99.60	75.00	41.70	26.10	99.40

(a)

Watermarker	KGW$_{weak}$				KGW$_{strong}$			
Paraphraser	SIR$_{weak}$		SIR$_{strong}$		SIR$_{weak}$		SIR$_{strong}$	
FPR	D_W	D_P	D_W	D_P	D_W	D_P	D_W	D_P
1%	3.40	90.09	0.20	92.77	9.00	79.26	0.80	90.52
5%	5.90	92.12	2.00	94.66	14.10	89.41	2.50	93.76
10%	7.80	93.81	3.80	95.88	17.30	92.14	4.70	95.88

(b)

Watermarker	KGW$_{weak}$				KGW$_{strong}$			
Paraphraser	PRW$_{weak}$		PRW$_{strong}$		PRW$_{weak}$		PRW$_{strong}$	
FPR	D_W	D_P	D_W	D_P	D_W	D_P	D_W	D_P
1%	41.10	48.00	6.20	99.90	56.00	44.40	14.80	99.70
5%	54.40	65.70	16.70	99.90	64.70	64.30	25.40	100.00
10%	60.20	71.70	23.70	99.90	68.50	70.80	33.00	100.00

(c)

Watermarker	SIR$_{weak}$				SIR$_{strong}$			
Paraphraser	KGW$_{weak}$		KGW$_{strong}$		KGW$_{weak}$		KGW$_{strong}$	
FPR	D_W	D_P	D_W	D_P	D_W	D_P	D_W	D_P
1%	55.74	25.10	12.45	96.70	61.55	21.00	10.98	97.40
5%	65.97	34.80	25.86	99.20	72.95	30.90	27.56	98.80
10%	69.73	40.30	31.22	99.50	75.44	38.70	34.23	99.40

(d)

Watermarker	SIR$_{weak}$				SIR$_{strong}$			
Paraphraser	PRW$_{weak}$		PRW$_{strong}$		PRW$_{weak}$		PRW$_{strong}$	
FPR	D_W	D_P	D_W	D_P	D_W	D_P	D_W	D_P
1%	41.05	19.80	13.22	97.50	45.58	14.00	12.58	97.10
5%	52.67	56.60	22.43	100.00	59.98	45.90	25.48	99.70
10%	56.17	67.30	27.63	100.00	63.17	59.00	30.49	99.90

(e)

Watermarker	PRW$_{weak}$				PRW$_{strong}$			
Paraphraser	KGW$_{weak}$		KGW$_{strong}$		KGW$_{weak}$		KGW$_{strong}$	
FPR	D_W	D_P	D_W	D_P	D_W	D_P	D_W	D_P
1%	37.00	28.20	14.60	96.70	67.90	18.30	14.60	96.70
5%	56.40	38.60	22.30	99.20	73.50	29.70	22.30	99.20
10%	65.80	45.70	30.20	99.40	76.50	38.70	30.20	99.40

(f)

Watermarker	PRW$_{weak}$				PRW$_{strong}$			
Paraphraser	SIR$_{weak}$		SIR$_{strong}$		SIR$_{weak}$		SIR$_{strong}$	
FPR	D_W	D_P	D_W	D_P	D_W	D_P	D_W	D_P
1%	22.30	79.56	29.40	91.91	32.00	65.31	37.10	79.00
5%	31.80	91.00	37.70	94.43	42.20	87.78	47.10	90.89
10%	39.20	92.56	42.10	95.63	47.90	91.19	51.20	93.22

(g)

Watermarker	PRW$_{weak}$				PRW$_{strong}$			
Paraphraser	PRW$_{weak}$		PRW$_{strong}$		PRW$_{weak}$		PRW$_{strong}$	
FPR	D_W	D_P	D_W	D_P	D_W	D_P	D_W	D_P
1%	26.60	41.20	9.30	99.90	44.20	41.70	11.20	99.30
5%	37.80	67.20	14.30	100.00	50.60	58.50	18.60	99.90
10%	45.30	72.00	18.60	100.00	55.50	65.20	24.90	99.90

(h)

Table 6: TPR of the paraphrased text T_P with dual watermarks when utilizing LLaMA-2-13B as the base model of watermarker. FPR is set to 1%, 2% & 5%, respectively.

Watermarker	KGW$_{weak}$				KGW$_{strong}$			
Paraphraser	KGW$_{weak}$		KGW$_{strong}$		KGW$_{weak}$		KGW$_{strong}$	
FPR	D_W	D_P	D_W	D_P	D_W	D_P	D_W	D_P
1%	1.40	23.40	4.30	98.00	1.50	26.30	6.90	97.90
5%	10.90	34.30	17.60	99.40	7.30	36.50	19.90	99.20
10%	17.20	44.10	25.60	99.60	15.30	44.20	28.30	99.40

(a)

Watermarker	KGW$_{weak}$				KGW$_{strong}$			
Paraphraser	SIR$_{weak}$		SIR$_{strong}$		SIR$_{weak}$		SIR$_{strong}$	
FPR	D_W	D_P	D_W	D_P	D_W	D_P	D_W	D_P
1%	0.80	92.81	0.90	85.32	0.90	72.27	1.20	87.80
5%	3.60	94.72	3.80	91.24	3.80	87.62	4.10	93.63
10%	6.70	96.18	6.70	93.65	7.40	91.60	8.10	95.49

(b)

Watermarker	KGW$_{weak}$				KGW$_{strong}$			
Paraphraser	PRW$_{weak}$		PRW$_{strong}$		PRW$_{weak}$		PRW$_{strong}$	
FPR	D_W	D_P	D_W	D_P	D_W	D_P	D_W	D_P
1%	5.00	36.80	6.90	99.80	5.80	14.70	8.80	97.40
5%	21.10	56.10	22.30	100.00	17.80	52.30	22.00	99.80
10%	34.30	66.20	31.60	100.00	29.00	63.00	31.50	100.00

(c)

Watermarker	SIR$_{weak}$				SIR$_{strong}$			
Paraphraser	KGW$_{weak}$		KGW$_{strong}$		KGW$_{weak}$		KGW$_{strong}$	
FPR	D_W	D_P	D_W	D_P	D_W	D_P	D_W	D_P
1%	1.25	19.30	5.17	96.40	2.61	17.00	6.52	93.70
5%	4.05	32.30	15.82	98.30	6.58	26.90	16.04	98.80
10%	6.76	40.50	20.34	98.70	10.44	35.20	20.32	99.30

(d)

Watermarker	PRW$_{weak}$				PRW$_{strong}$			
Paraphraser	PRW$_{weak}$		PRW$_{strong}$		PRW$_{weak}$		PRW$_{strong}$	
FPR	D_W	D_P	D_W	D_P	D_W	D_P	D_W	D_P
1%	2.52	27.20	4.26	97.70	3.76	13.70	8.25	97.80
5%	5.45	52.20	10.93	99.70	8.57	46.00	13.14	100.00
10%	8.28	59.70	13.77	99.80	11.60	55.60	15.53	100.00

(e)

Watermarker	PRW$_{weak}$				PRW$_{strong}$			
Paraphraser	KGW$_{weak}$		KGW$_{strong}$		KGW$_{weak}$		KGW$_{strong}$	
FPR	D_W	D_P	D_W	D_P	D_W	D_P	D_W	D_P
1%	0.10	23.50	5.70	98.60	1.30	17.50	7.80	96.80
5%	2.90	35.60	17.30	99.40	6.30	29.70	21.80	98.60
10%	13.60	44.20	34.00	99.50	11.70	42.10	29.60	98.70

(f)

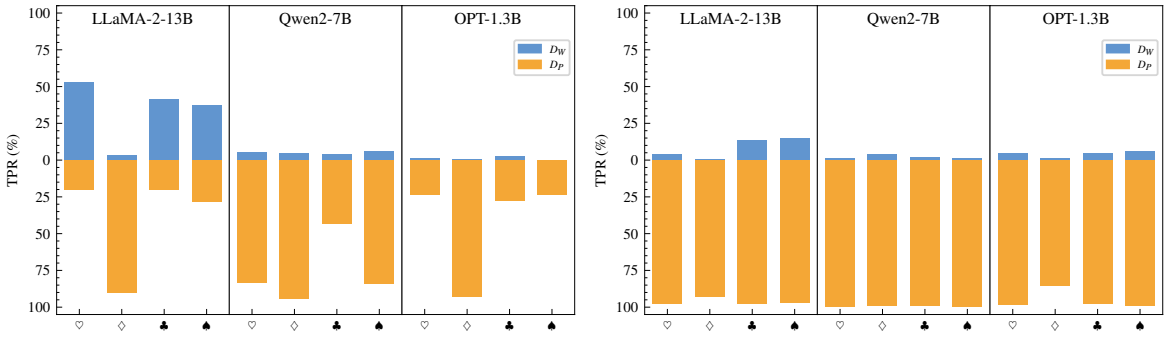
Watermarker	PRW$_{weak}$				PRW$_{strong}$			
Paraphraser	SIR$_{weak}$		SIR$_{strong}$		SIR$_{weak}$		SIR$_{strong}$	
FPR	D_W	D_P	D_W	D_P	D_W	D_P	D_W	D_P
1%	10.60	87.23	24.90	88.96	14.00	78.84	24.30	88.62
5%	24.30	91.38	38.60	93.93	30.30	89.81	38.50	93.53
10%	38.90	92.50	48.60	94.92	36.60	92.39	43.60	94.53

(g)

Watermarker	PRW$_{weak}$				PRW$_{strong}$			
Paraphraser	PRW$_{weak}$		PRW$_{strong}$		PRW$_{weak}$		PRW$_{strong}$	
FPR	D_W	D_P	D_W	D_P	D_W	D_P	D_W	D_P
1%	0.80	36.10	4.70	99.60	1.00	11.60	6.10	93.30
5%	3.70	57.80	10.70	100.00	6.50	52.10	13.80	100.00
10%	10.30	66.00	19.90	100.00	11.60	63.20	18.40	100.00

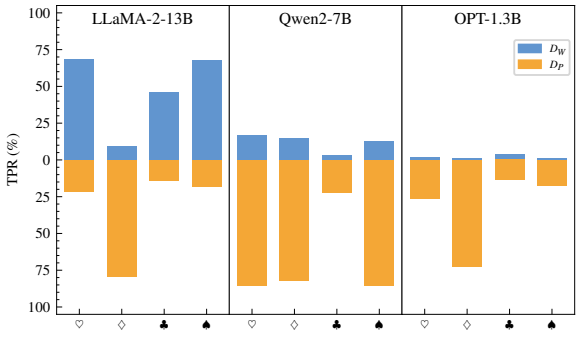
(h)

Table 7: TPR of the paraphrased text T_P with dual watermarks when utilizing OPT-1.3B as the base model of watermarker. FPR is set to 1%, 2% & 5%, respectively.

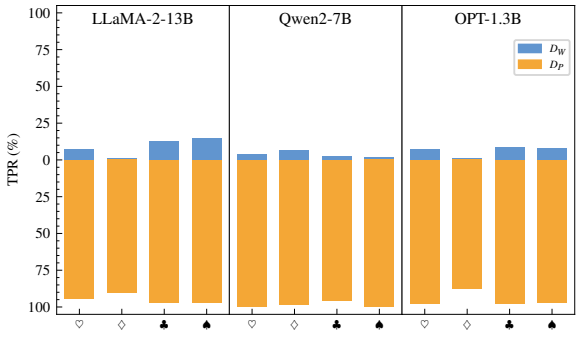


(a) weak W, weak P

(b) weak W, strong P



(c) strong W, weak P



(d) strong W, strong P

Figure 6: TPR of the *paraphrased* text T_P with different settings and base models as both the watermarker and paraphraser. Blue bars represent TPRs detected by the original watermark detector D_W, and orange bars are TPRs detected by the collider D_P. Symbols ♡, ♢, ♣, ♠ denote different watermarker-paraphraser pairs as KGW-KGW, KGW-SIR, SIR-PRW and PRW-KGW, respectively.

```
[INST] «SYS»
Assume you are a helpful assistant.
Your job is to translate the given
text from LANGUAGE to LANGUAGE.
«/SYS»
{INPUT_TEXT} [/INST]
You're welcome!  Here's a translated
version of the original text:
```

Figure 7: The back-translation prompt template for LLaMA-2 translator.

W \ P	P'	KGW$_{\texttt{wk}}$	PRW$_{\texttt{wk}}$	SIR$_{\texttt{wk}}$
KGW$_{\texttt{wk}}$	0.879	0.852	0.858	0.790
PRW$_{\texttt{wk}}$	0.888	0.864	0.872	0.781
SIR$_{\texttt{wk}}$	0.867	0.852	0.873	-

(a)

W \ P	P'	KGW$_{\texttt{wk}}$	PRW$_{\texttt{wk}}$	SIR$_{\texttt{wk}}$
KGW$_{\texttt{wk}}$	0.879	0.714	0.722	0.699
PRW$_{\texttt{wk}}$	0.888	0.721	0.726	0.709
SIR$_{\texttt{wk}}$	0.867	0.730	0.748	-

(b)

W \ P	P'	KGW$_{\texttt{wk}}$	PRW$_{\texttt{wk}}$	SIR$_{\texttt{wk}}$
KGW$_{\texttt{sg}}$	0.874	0.847	0.857	0.788
PRW$_{\texttt{sg}}$	0.887	0.849	0.867	0.776
SIR$_{\texttt{sg}}$	0.883	0.851	0.868	-

(c)

W \ P	P'	KGW$_{\texttt{sg}}$	PRW$_{\texttt{sg}}$	SIR$_{\texttt{sg}}$
KGW$_{\texttt{sg}}$	0.874	0.706	0.725	0.696
PRW$_{\texttt{sg}}$	0.887	0.719	0.721	0.689
SIR$_{\texttt{sg}}$	0.883	0.721	0.737	-

(d)

Table 8: Semantic similarity between paraphrased text and original text of LLaMA-2-13B.

Watermarker			KGW$_{wk}$											
Translator	No Watermark		KGW$_{wk}$		KGW$_{sg}$		PRW$_{wk}$		PRW$_{sg}$		SIR$_{wk}$		SIR$_{sg}$	
FPR	Gen TPR	D_W	Old TPR	New TPR	Old TPR	New TPR	Old TPR	New TPR	Old TPR	New TPR	Old TPR	New TPR	Old TPR	New TPR
1%	99.9	44.8	41.9	9.2	28.5	69.5	34.2	13.7	18.5	58.5	31.4	5.15	3.9	93.5
5%	100	69.7	66.5	19.6	51.2	79.1	59.8	35.8	40.7	82.8	56.8	30.37	14.6	95.56
10%	100	78.4	75.2	30.3	63	85.5	70.3	50.1	55	89.9	68.4	46.92	21.8	96.08

(a)

Watermarker			KGW$_{sg}$											
Translator	No Watermark		KGW$_{wk}$		KGW$_{sg}$		PRW$_{wk}$		PRW$_{sg}$		SIR$_{wk}$		SIR$_{sg}$	
FPR	Gen TPR	D_W	Old TPR	New TPR	Old TPR	New TPR	Old TPR	New TPR	Old TPR	New TPR	Old TPR	New TPR	Old TPR	New TPR
1%	100	69.3	67.8	8.1	50	68.9	61.9	13.6	38.8	53	55.3	4.09	7.5	91.7
5%	100	80.7	80.1	25.3	66.8	83	76	40.6	56.7	84.3	70.8	36.09	20.8	95.95
10%	100	86.4	84.1	35.2	74.5	88.2	81.8	58.4	65.9	92.5	76.9	52.25	29.6	96.89

(b)

Watermarker			PRW$_{wk}$											
Translator	No Watermark		KGW$_{wk}$		KGW$_{sg}$		PRW$_{wk}$		PRW$_{sg}$		SIR$_{wk}$		SIR$_{sg}$	
FPR	Gen TPR	D_W	Old TPR	New TPR	Old TPR	New TPR	Old TPR	New TPR	Old TPR	New TPR	Old TPR	New TPR	Old TPR	New TPR
1%	95.4	32.9	**34.1**	7.5	23	68	21.8	12.7	14.3	54.6	25.2	5.73	10.8	92.87
5%	100	59.3	**61.1**	20.6	46	81	45.4	36	31.1	78.9	49.1	25.86	18.7	95.56
10%	100	78	77.7	32	65.1	86.3	67.3	51.7	50.5	88.4	69.1	44.67	31.3	96.49

(c)

Watermarker			PRW$_{sg}$											
Translator	No Watermark		KGW$_{wk}$		KGW$_{sg}$		PRW$_{wk}$		PRW$_{sg}$		SIR$_{wk}$		SIR$_{sg}$	
FPR	Gen TPR	D_W	Old TPR	New TPR	Old TPR	New TPR	Old TPR	New TPR	Old TPR	New TPR	Old TPR	New TPR	Old TPR	New TPR
1%	99.7	63.9	**64.8**	9.6	50.1	67.9	55.2	10.2	36.6	42.4	50.6	3.56	20.1	90.44
5%	100	80.2	**81.1**	22.5	71.2	81	76.9	32.2	57.7	75	72.8	27.7	33.6	94.49
10%	100	87	86.8	33.7	80.5	87	83.4	52.8	68.3	89	80.7	46.95	45.1	95.43

(d)

Watermarker			SIR$_{wk}$											
Translator	No Watermark		KGW$_{wk}$		KGW$_{sg}$		PRW$_{wk}$		PRW$_{sg}$		SIR$_{wk}$		SIR$_{sg}$	
FPR	Gen TPR	D_W	Old TPR	New TPR	Old TPR	New TPR	Old TPR	New TPR	Old TPR	New TPR	Old TPR	New TPR	Old TPR	New TPR
1%	87.9	5.6	**4.75**	7.7	4.28	67.4	**6.12**	11.5	4.4	41.6	-	-	-	-
5%	95.3	18.1	**14.26**	24.7	16.9	84	**22.67**	31	17.3	72.6	-	-	-	-
10%	96.6	26.3	21.74	37.7	24.75	89.8	**32.4**	52.8	25.7	86.9	-	-	-	-

(e)

Watermarker			SIR$_{sg}$											
Translator	No Watermark		KGW$_{wk}$		KGW$_{sg}$		PRW$_{wk}$		PRW$_{sg}$		SIR$_{wk}$		SIR$_{sg}$	
FPR	Gen TPR	D_W	Old TPR	New TPR	Old TPR	New TPR	Old TPR	New TPR	Old TPR	New TPR	Old TPR	New TPR	Old TPR	New TPR
1%	94.3	3	**1.74**	9.4	**4.15**	74.4	**3.82**	6.7	2.51	22.5	-	-	-	-
5%	97.8	16.2	**14.23**	30.4	**17.2**	86.9	**19.98**	34.4	15.15	73	-	-	-	-
10%	98.3	24.5	23.75	42.2	**27.98**	91.2	**28.71**	51.5	23.87	84.5	-	-	-	-

(f)

Table 9: TPR of the back-translation text T_P with dual watermarks when utilizing LLaMA-2-13B as the base model of watermarker. FPR is set to 1%, 2% & 5%, respectively.

Watermarker	KGW$_{wk}$													
Translator	No Watermark		KGW$_{wk}$		KGW$_{sg}$		PRW$_{wk}$		PRW$_{sg}$		SIR$_{wk}$		SIR$_{sg}$	
FPR	Gen TPR	D_W	Old TPR	New TPR	Old TPR	New TPR	Old TPR	New TPR	Old TPR	New TPR	Old TPR	New TPR	Old TPR	New TPR
1%	99.9	6	5.4	83.3	1.2	99.8	2.8	55.8	0.3	99.4	4.5	94.15	3.8	99
5%	100	18.8	16.6	93	5.7	99.8	10.1	87.4	2.1	99.9	16.7	97.88	14.7	99.4
10%	100	28.2	25.3	96.5	13.1	99.8	17.9	94.1	4.3	99.9	29.2	98.39	25.9	99.8

(a)

Watermarker	KGW$_{sg}$													
Translator	No Watermark		KGW$_{wk}$		KGW$_{sg}$		PRW$_{wk}$		PRW$_{sg}$		SIR$_{wk}$		SIR$_{sg}$	
FPR	Gen TPR	D_W	Old TPR	New TPR	Old TPR	New TPR	Old TPR	New TPR	Old TPR	New TPR	Old TPR	New TPR	Old TPR	New TPR
1%	100	19.2	16.6	85.1	3.6	99.9	12	48	2.1	98.9	14.5	82.09	6.5	98.29
5%	100	35	32.5	93.4	12.3	99.9	25.9	88.2	6.5	100	32.5	97.03	20.3	99.8
10%	100	46.8	45	95.3	20.5	99.9	34.5	94.4	11.5	100	45.3	97.95	31.8	100

(b)

Watermarker	PRW$_{wk}$													
Translator	No Watermark		KGW$_{wk}$		KGW$_{sg}$		PRW$_{wk}$		PRW$_{sg}$		SIR$_{wk}$		SIR$_{sg}$	
FPR	Gen TPR	D_W	Old TPR	New TPR	Old TPR	New TPR	Old TPR	New TPR	Old TPR	New TPR	Old TPR	New TPR	Old TPR	New TPR
1%	95.4	8.7	5.8	84.2	1.5	99.8	6.6	46.2	2.8	99.5	5.4	94.47	2.3	99.3
5%	100	20.4	16.4	91.4	4.5	99.9	15.8	81.5	9.2	99.8	15.9	97.69	5.9	99.7
10%	100	39.4	29.4	94.9	8.1	99.9	33.3	91.1	21.5	99.8	31.7	98.69	18.9	99.7

(c)

Watermarker	PRW$_{sg}$													
Translator	No Watermark		KGW$_{wk}$		KGW$_{sg}$		PRW$_{wk}$		PRW$_{sg}$		SIR$_{wk}$		SIR$_{sg}$	
FPR	Gen TPR	D_W	Old TPR	New TPR	Old TPR	New TPR	Old TPR	New TPR	Old TPR	New TPR	Old TPR	New TPR	Old TPR	New TPR
1%	99.7	18.7	12.7	85.4	1.4	100	13.1	28.8	6.9	97.2	12.9	86.77	16.3	99.3
5%	100	43.1	33.3	93	7.5	100	37.4	82.1	20.4	99.7	36.6	95.39	39	99.7
10%	100	58.1	48.6	96.6	13.8	100	54	91.5	35	100	55.4	97.7	57.8	99.7

(d)

Watermarker	SIR$_{wk}$													
Translator	No Watermark		KGW$_{wk}$		KGW$_{sg}$		PRW$_{wk}$		PRW$_{sg}$		SIR$_{wk}$		SIR$_{sg}$	
FPR	Gen TPR	D_W	Old TPR	New TPR	Old TPR	New TPR	Old TPR	New TPR	Old TPR	New TPR	Old TPR	New TPR	Old TPR	New TPR
1%	87.9	3.1	2.86	72	0.8	99.9	3.89	43.1	1.52	98.8	-	-	-	-
5%	95.3	11.2	11.03	86.2	4.81	100	10.84	77.8	6.07	100	-	-	-	-
10%	96.6	18.6	16.96	93.4	10.03	100	16.36	88.4	11.23	100	-	-	-	-

(e)

Watermarker	SIR$_{sg}$													
Translator	No Watermark		KGW$_{wk}$		KGW$_{sg}$		PRW$_{wk}$		PRW$_{sg}$		SIR$_{wk}$		SIR$_{sg}$	
FPR	Gen TPR	D_W	Old TPR	New TPR	Old TPR	New TPR	Old TPR	New TPR	Old TPR	New TPR	Old TPR	New TPR	Old TPR	New TPR
1%	94.3	2.5	2.28	77.6	0.81	100	2.93	22.4	2.63	95.4	-	-	-	-
5%	97.8	10.7	12.22	89.4	5.85	100	11.83	75.6	7.48	99.8	-	-	-	-
10%	98.3	18.6	21.43	94.4	12.11	100	20.52	88.2	13.04	100	-	-	-	-

(f)

Table 10: TPR of the paraphrased text T_P with dual watermarks when utilizing QWEN-7B as the base model of watermarker. FPR is set to 1%, 2% & 5%, respectively.

W \ P	$\varnothing$	P'	KGW_{wk}	PRW_{wk}	SIR_{wk}
KGW_{wk}	in the future of our schools. Voters approved $700 million for school construction and improvement projects, a new building program and the sale of a new $70 million bond. The new bond will be for projects to improve the air quality in our schools and make the buildings safer and more secure. We are excited to share the following updates on the progress of the bond projects. We have scheduled all projects through 2018 and are beginning to schedule some projects for the 2019 school year. We are continuing to work with the District's Capital Planning Committee to	In the future of our schools, voters approved $700 million for school construction and improvement projects, a new building program and the sale of a new $70 million bond. The new bond will be for projects to improve the air quality in our schools and make the buildings safer and more secure. We are excited to share the following updates on the progress of the bond projects. We have scheduled all projects through 2018 and are beginning to schedule some projects for the 2019 school year. We are continuing to work with the District's Capital Planning Committee to	in the future of our schools. Voters approved $700 million for school construction and improvement projects, a new building program and the sale of a new $70 million bond. The new bond will be for projects to improve the air quality in our schools and make the buildings safer and more secure. We are excited to share the following updates on the progress of the bond projects. We have scheduled all projects through 2018 and are beginning to schedule some projects for the 2019 school year. We are continuing to work with the District's Capital Planning Committee to	In the future, I hope that our school district will be able to make a lot of changes, like building new school buildings and making our schools more safe and secure. Voters voted to let our school district sell $70,000,000.00 of bonds for rebuilding and renovating our schools. The reason why I decided to write this letter is because I want to make a change and help with our school district. We need to update the school buildings and make our schools safe and secure.	in our schools' futures. Approximately 700 million dollars were approved by voters for school construction and improvement projects, a new building program, and the sale of a new bond worth about 70 million dollars. The new bond will be used for initiatives that will make our schools' facilities safer and more secure as well as improve the air quality. We are thrilled to provide the following updates on the bond projects' status. All of the projects are scheduled through 2018, and we are starting to
PRW_{wk}	Jessie said, "I don't see her being a rude, bad person. What a lot of these women do when you see a lot of these women being rude, and you see a lot of these women not being respectful, and a lot of these women not being sweet, and a lot of these women being difficult,	Jessie said, "I don't see her being a rude, bad person. What a lot of these women do when you see a lot of these women being rude, and you see a lot of these women not being respectful, and a lot of these women not being sweet, and a lot of these women being difficult,	Jessie said, "I don't see her being a rude, bad person. What a lot of these women do when you see a lot of these women being rude, and you see a lot of these women being not respectful, and a lot of these women being not sweet, and a lot of these women being difficult, what you see when you see these women being difficult,	Jessie said, "I don't see her being a rude, bad person. What a lot of these women do when you see a lot of these women being rude, and you see a lot of these women being respectful, and a lot of these women not being respectful, and a lot of these women not being sweet, and a lot of these women being difficult,	Jessie said, I don't see her being a rude, bad person, but what I see is a lot of women being rude, and what I see is a lot of women not being respectful, and what I see is a lot of women not being sweet and difficult,
SIR_{wk}	moving into agency work in 2008 and subsequently working on a freelance, interim and consultancy basis before re-joining Workhouse in 2018, working on a freelance contractor basis alongside a day job in a corporate agency environment and a project-based assignment in a digital transformation project in aerospace and defence, and a project-based assignment in aerospace and defence, before deciding to join on a permanent capacity and take on a client account and agency-wide responsibility for a leading client, which is currently in development and will be launched in	moving into agency work in 2008 and subsequently working on a freelance, interim and consultancy basis before re-joining Workhouse in 2018, working on a freelance contractor basis alongside a day job in a corporate agency environment and a project-based assignment in aerospace and defence, and a project-based assignment in aerospace and defence, before deciding to join on a permanent capacity and take on a client account and agency-wide responsibility for a leading client, which is currently in development and will be launched in	moving into agency work in 2008 and subsequently working on a freelance, interim and consultancy basis before re-joining Workhouse in 2018, working on a freelance contractor basis alongside a day job in a corporate agency environment and a project-based assignment in aerospace and defence, and a project-based assignment in aerospace and defence, before deciding to join on	moving into agency work in 2008 and subsequently working on a freelance, interim and consultancy basis before re-joining Workhouse in 2018, working on a freelance contractor basis alongside a day job in a corporate agency environment and a project-based assignment in aerospace and defence, and a project-based assignment in aerospace and defence, before deciding to join on a permanent capacity and take on a client account and agency-wide responsibility for a leading client, which is currently in development and will be launched in	N/A

Table 11: Examples of watermarked texts under **weak** W, **weak** P settings.

W \ P	$\varnothing$	P'	KGW_{wk}	PRW_{wk}	SIR_{wk}
KGW_{wk}	the past, since the portable battery pack from ZMI provides mobile power so your mobile devices never run out again. You will never be out of the loop again. Whether at a business appointment in the city, a vacation in the country, a hike in the high alps - the ZMI portable batteries provide mobile power supply, so your mobile devices never let you down again. Choosing ZMI mobile batteries means investing money sensibly.	The ZMI portable power station provides mobile power so your mobile devices never run out again. You'll never be out of the loop again. Whether you're at a business meeting in the city, on vacation in the country, or hiking in the high alps - the ZMI portable batteries provide mobile power supply, so your mobile devices never let you down again. By choosing ZMI mobile batteries, you're investing money sensibly.	In the past, the portable power bank from ZMI was a lifesaver when your mobile devices were running out of power, so you never had to be out of the loop again. No matter where you were, whether in the city on a business trip, on vacation in the countryside, hiking in the high alps - the ZMI power bank always provided the power to keep your mobile devices up and running. When you buy	In the past, since the battery from ZMI provides mobile power, your mobile devices never run out again. You will never be out of the loop again. Whether at a business appointment in the city, a vacation in the country, a hike in the high alps - the ZMI batteries provide mobile power supply, so your mobile devices never let you down again. Choosing ZMI mobile batteries means investing money sensibly.	In the past, your mobile devices never ran out of power again since ZMI's battery supplied mobile power. Never again will you be left behind. ZMI batteries ensure that your mobile devices never fail you, whether you're on a work trip in the city, on holiday in the country, or hiking in the high alps. Selecting ZMI mobile batteries is a wise financial decision.
PRW_{wk}	ot seeks a full-time Director, Marketing and Communications. As a member and leader of the organization, the Director, Marketing and Communications will play a crucially important role and have the ability and skills needed for the organization and the community. With the recent completion and reopening of the historic theater, the community can look ahead and the new Director, Marketing and Communications can help define and execute a long-standing strategic marketing and communications strategy.	The community seeks a full-time Director, Marketing and Communications. As a member and leader of the organization, the Director, Marketing and Communications will play a crucially important role and have the ability and skills needed for the organization and the community. With the recent completion and reopening of the historic theater, the community can look ahead and the new Director, Marketing and Communications can help define and execute a long-standing strateg	A full-time Director of Marketing and Communications is needed for the community. The Director of Marketing and Communications will be a key member and leader of the organization, possessing the abilities and talents required for the community and the organization. The community can turn to the future now that the historic theater has been completed and reopened, and the new director of marketing and communications can assist in defining and carrying out a long-standing strategy.	Our organization seeks a new, forward-facing, and creative assistant with a varied skillset and an eagerness to jump headfirst and learn from our expert staff members, who have more combined years' experience with our organization and its constituents' needs and goals, and our overall vision—the most important, but sometimes difficult, task—which is that everyone, whether or not our organization is directly related or connected, is a member, and our organization's members and supporters are our most important asset	Our organization is looking for a new, creative, forward-thinking assistant with a variety of skills and a willingness to dive right in and learn from our knowledgeable staff members who have more combined years of experience with our organization, the needs and goals of our constituents, and our overall vision—the most crucial, but occasionally challenging, task—which is that everyone is a member, whether or not our organization is directly related or connected, and that our members and supporters are our most valuable,
SIR_{wk}	lending institutions and capital partners, as well as prospected new clientele and their professional advisors in their evaluative process of our client'ss financial information and capabilities, and its presentation of their financial and operational health and well-being, and their potential success and growth in their respective industry and region of operations and beyond, as well as their readiedness and ability in their ability and preparedness	Lending institutions and capital partners, as well as prospected new clientele and their professional advisors in their evaluative process of our client'ss financial information and capabilities, as well as their presentation of their financial and operational health and well- being, and their potential success and growth in their respective industry and region of operations and beyond, as well as their readiedness and ability in their ability and preparedness	during their audit and due diligence process, when our clients are looking for new lines and higher limits, as well as their presentation and representation of their financial and operations health and well, their potential and actual increase and overall value, as well as their readiness, suitability, and overall availability, we prospected new clients and their financial and business advisors	lending institutions and capital partners, as well as prospected new clients and their financial and business advisors, during their audit and due diligence process, when our clients are seeking new lines and higher limits, and their presentation and representation of their financial and operations health and well, and their potential and actual increase and overall value, and their preparedness and suitability and overall availabilty,	N/A

Table 12: Examples of watermarked texts under **strong** W, **strong** P settings.