

Watermarking Large Language Models: An Unbiased and Low-risk Method

Minjia Mao^{1*} Dongjun Wei² Zeyu Chen¹ Xiao Fang¹ Michael Chau²

¹University of Delaware

²The University of Hong Kong

{mjmao, chenze, xfang}@udel.edu

dongjun@connect.hku.hk, mchau@business.hku.hk


Abstract


Recent advancements in large language models (LLMs) have highlighted the risk of misusing them, raising the need for accurate detection of LLM-generated content. In response, a viable solution is to inject imperceptible identifiers into LLMs, known as watermarks. Our research extends the existing watermarking methods by proposing the novel Sampling One Then Accepting (STA-1) method. STA-1 is an unbiased watermark that preserves the original token distribution in expectation and has a lower risk of producing unsatisfactory outputs in low-entropy scenarios compared to existing unbiased watermarks. In watermark detection, STA-1 does not require prompts or a white-box LLM, provides statistical guarantees, demonstrates high efficiency in detection time, and remains robust against various watermarking attacks. Experimental results on low-entropy and high-entropy datasets demonstrate that STA-1 achieves the above properties simultaneously, making it a desirable solution for watermarking LLMs. Implementation codes for this study are available online.¹

1 Introduction


Large language models (LLMs) are large-scale deep learning models that can understand and generate natural languages by learning from a large amount of textual data. As LLMs can generate content more efficiently at a lower cost compared to humans, the risk of LLMs being employed to generate biased, fake, or malicious content is also increasing (Mirsky et al., 2023; Fang et al., 2024; Pan et al., 2023). To reduce the harm caused by LLMs, it is crucial to identify LLM-generated content precisely and efficiently (Kirchenbauer et al., 2023b). A viable solution is to inject watermarks


into LLM-generated text. The watermarked text is imperceptible to humans but detectable by certain models (Liu et al., 2023b). This is achieved by controlling the randomness of the token generation process in LLMs (Kirchenbauer et al., 2023a; Lee et al., 2023), with the randomness kept confidential by LLM owners. In this study, we seek a watermark with the following properties during the generation phase, which are crucial for an effective watermark:


 **Unbiased:** The watermark should adjust the probability distribution while maintaining the same expectation as the unwatermarked distribution, making it impossible to discern between watermarked and unwatermarked text.


 **Low-risk:** The watermark should have a low risk of producing unsatisfactory outputs in *low-entropy* scenarios (e.g., code generation), where high-probability tokens should be sampled even with watermarks.

Furthermore, the watermark should have the following necessary properties during the detection phase:

 **Black-box:** We do not need prompts or a white-box LLM for detection.

 **Guarantee:** We can have a statistical guarantee on the type II error, where the watermark detection fails to identify a watermarked text.

 **Efficiency:** The detection should only require a low time complexity.

 **Robustness:** The watermark is hard to be removed by watermarking attacks.

However, we find that existing watermarking methods cannot satisfy all these properties simultaneously in the generation and detection phases. In response to these challenges, we propose a novel Sampling One Then Accepting (STA-1) method that can simultaneously achieve these properties. We provide an analysis of previous methods in Appendix A and compare them with the proposed

*Minjia Mao, Dongjun Wei, and Zeyu Chen contribute equally. Corresponding to Minjia Mao.

¹<https://github.com/djwei96/STA>

Table 1: Comparison of the Properties of the Proposed Watermark with Properties of Previous Methods.

Method	Watermark Generation		Watermark Detection			
	⚖️ Unbiased	🌈 Low-risk	✅ Guarantee	🚫 Black-box	⚡ Efficiency	🛡️ Robustness
Kirchenbauer et al. (2023a)			✓	✓	✓	✓
Lee et al. (2023)				✓	✓	✓
Hu et al. (2024)	✓				✓	
Christ et al. (2023)	✓		✓	✓	✓	
Kuditipudi et al. (2023)	✓			✓		✓
Lu et al. (2024a)			✓	✓	✓	✓
Wu et al. (2024)	✓			✓	✓	✓
Dathathri et al. (2024)	✓			✓	✓	
Ours (STA-1)	✓	✓	✓	✓	✓	✓

STA-1 method in Table 1. Our proposed STA-1 method can be traced back to the original watermarking method (denoted as KGW) (Kirchenbauer et al., 2023a), where the token set is divided into a green and a red list at each generation step. Instead of raising logits in the green list, STA-1 samples a token from the original probability distribution and accepts it if it is in the green list. If the sampled token is in the red list, it resamples another token and accepts it. We theoretically prove that our STA-1 method is an unbiased watermarking method, which is similar to previous unbiased watermarks (Hu et al., 2024; Wu et al., 2024).

The STA-1 method also outperforms other unbiased watermarks in low-entropy scenarios with a lower risk of producing unsatisfactory outputs. Specifically, unsatisfactory outputs in low-entropy scenarios represent that under certain watermark keys, the unbiased watermarking method alters the probability distribution too much such that high-probability tokens cannot be sampled at risk. We prove that STA-1 is less risky than previous unbiased watermarks by analyzing the variance of the probability after altering, using a well-adopted risk-return analysis (Sharpe, 1998).

Another benefit of our proposed method is that STA-1 is a natural extension of KGW that inherits its advantages in the detection phase. Specifically, STA-1 counts the number of green list tokens and employs the z -test for watermark detection. The z -test naturally eliminates the need for prompts and white-box LLMs in detection (which is required in some previous work (Hu et al., 2024)) and only requires $O(m)$ time complexity, where m is the number of tokens. Furthermore, we establish the statistical guarantees for the type II error in watermark detection. These guarantees are related to the Gini index of the probability distribution, a com-

mon metric in machine learning (Breiman, 2017), compared to the proposed Spike entropy in KGW.

Additionally, we propose STA-M, an extension of STA-1, by setting up a threshold for entropy in generation (Lee et al., 2023; Wang et al., 2023) and sampling more times for high-entropy steps. Although STA-M is not unbiased theoretically, it allows higher watermark strength with small text quality shifts empirically. Based on the experimental results, we also find that our proposed STA-M method has better robustness compared to KGW against various attacks. Our main contributions can be summarized as follows:

1. We propose STA-1, a novel unbiased watermarking method that has a lower risk theoretically compared to other unbiased watermarks. Moreover, we introduce STA-M, an extension of STA-1 that enhances watermark strength with low text quality shifts.
2. We prove that STA-1 has statistical guarantees for the type II error in its detection based on the widely used Gini index. STA-1 also does not require access to white-box LLMs and only requires $O(m)$ time complexity in detection.
3. Experimental results on low-entropy and high-entropy datasets empirically show that STA-1 is unbiased and has a lower risk of unsatisfactory outputs compared to other unbiased watermarks. Meanwhile, STA-M is more robust against different watermarking attacks than existing methods.

2 Preliminary

Notations. We follow notations in previous work (Kirchenbauer et al., 2023a; Hu et al., 2024) to

represent the generation task of LLMs. Let P_M denote a pretrained LLM and \mathcal{V} is the overall token set. An example token set contains more than 50,000 tokens ($|\mathcal{V}| > 50000$) (Radford et al., 2019). For simplicity, we use Python-style notation for an ordered token sequence, where $x^{-m:n} = (x^{-m}, x^{-m+1}, \dots, x^n)$, m and n are integers. In a typical LLM generation task, an LLM receives a sequence of $N_p + 1$ tokens $x^{-N_p:0}$, known as a prompt, and outputs a sequence of T tokens $x^{1:T}$ step by step. At step t , the probability of each token is given by the conditional distribution $P_M(x^t|x^{-N_p:(t-1)})$. The LLM generation follows an autoregressive fashion, where the joint probability of the generated tokens is $P_M(x^{1:T}|x^{-N_p:0}) = \prod_{t=1}^T P_M(x^t|x^{-N_p:(t-1)})$.

When applying watermarking methods, the LLM employs a private key k to adjust the conditional distribution from $P_M(x^t|x^{-N_p:(t-1)})$ to $P_{M,w}(x^t|x^{-N_p:(t-1)}; k)$, where $P_{M,w}$ indicates a watermarked LLM and the private key k is randomly selected from a key space K according to a known distribution $P_K(k)$. An unbiased watermark requires that the expectation of the watermarked distribution equals that of the original distribution (Hu et al., 2024), defined as follows.

Definition 1 (Unbiased watermark). Given a prompt $x^{-N_p:0}$ and a known distribution $P_K(k)$ of the key k , a watermarking method is unbiased towards the original model P_M if the watermarked model $P_{M,w}$ satisfies

$$\mathbb{E}_{k \sim P_K(k)} \left[P_{M,w}(x^t|x^{-N_p:(t-1)}; k) \right] = P_M(x^t|x^{-N_p:(t-1)}), \quad (1)$$

for any prompt $x^{-N_p:0} \in \mathcal{V}^{N_p+1}$, any token $x^t \in \mathcal{V}$, and all generation steps $1 \leq t \leq T$.

One of our main contributions is to show that the risk of unsatisfactory outputs in STA-1 is lower. Here, ‘risk’ is specifically defined for unbiased watermarks in low-entropy scenarios. To support our analysis, we introduce a previous biased watermark KGW (as the backbone of our study) (Kirchenbauer et al., 2023a), alongside other unbiased watermarks including Dipmark (Wu et al., 2024), γ -reweight (Hu et al., 2024), and RDW (Kuditipudi et al., 2023) in Appendix B.

3 Method: Sampling Then Accepting

In this section, we first propose the Sampling One Then Accepting (STA-1) method and theoretically

show that it is unbiased. We then analyze previous unbiased watermarks alongside STA-1 under a low-entropy protocol, showing that STA-1 has a lower risk of producing unsatisfactory outputs. Next, we explore the detection of STA-1-generated text using the z -test and provide a statistical guarantee for its type II error based on the Gini index. Finally, we introduce Sampling M Then Accepting (STA-M), an extension of STA-1.

3.1 Sampling One Then Accepting (STA-1)

We start by proposing the STA-1 method in Algorithm 1, which is always unbiased. We first utilize the last generated token from an LLM to compute its hash value and employ this value as the seed of a random number generator (RNG). We then use the RNG to divide the token set into a green and a red list (Kirchenbauer et al., 2023a). Finally, we sample from the original LLM output distribution (as depicted in Line 4 of Algorithm 1). If the token is in the green list (as shown in Lines 5 and 6 of Algorithm 1), we accept the sample. Otherwise, the token must be in the red list (as depicted in Lines 7 and 8 of Algorithm 1), and we sample a token again, always accepting the second sample.

Algorithm 1 STA-1 Text Generation

Input: A pretrained LLM P_M , a watermark key $k \in K$, the proportion of the green list $\gamma \in (0, 1)$, and a prompt $x^{-N_p:0}$

- 1: **for** $t = 1, 2, \dots, T$ **do**
- 2: Get the probability distribution of tokens $p^t = P_M(\cdot|x^{-N_p:(t-1)})$
- 3: Compute the hash of the last token x^{t-1} . Partition the token set \mathcal{V} to form the green G and red R lists based on the key k , the hash, and the proportion γ
- 4: Sample the candidate token x_c^t with p^t
- 5: **if** $x_c^t \in G$ **then**
- 6: **Accept the sampling**, the next generated token $x^t = x_c^t$
- 7: **else**
- 8: **Deny the sampling** (i.e., $x_c^t \in R$), sample x^t from the distribution p^t
- 9: **end if**
- 10: **end for**

Output: The generated text $x^{1:T}$

STA-1 is a simple yet effective method with many great properties. We begin by analyzing the unbiasedness of STA-1. In the following theorem, we assume that the key k is randomly sampled

from a uniform distribution. Consequently, the random partition of the green and red lists associated with this key is also uniform (Kirchenbauer et al., 2023a).

Theorem 1. *The STA-1 method (Algorithm 1) is an unbiased watermark.*

Proof. See Appendix C.1.

3.1.1 Risk in the Low-Entropy Scenario

The STA-1 method outperforms other unbiased watermarks in generating low-entropy texts, demonstrating a lower risk of producing unsatisfactory outputs. Specifically, the low-entropy text refers to a relatively deterministic sequence in natural language. The entropy measures the uncertainty of the probability distribution $P_M(x^t | x^{-N_p:(t-1)})$ at a single generation step among the token set \mathcal{V} , where low entropy means low uncertainty. For example, in code writing, the structure of a code sequence is regularized where few changes can be made (Lee et al., 2023). More explicitly, for a typical English pangram such as ‘The quick brown fox jumps over the lazy dog’ (Kirchenbauer et al., 2023a), both humans and machines should generate similar if not identical output. For example, when provided with the prompt ‘The quick brown fox jumps over the lazy’, the trained LLaMA-2-7B (Touvron et al., 2023) outputs an empirical probability above 0.8 for the next token ‘dog’. Such low-entropy scenarios are common in text generation tasks of LLMs. In this paper, we aim to model a simple problem protocol for the low-entropy generation scenario.

Low-entropy Protocol. For simplicity, we consider the low-entropy scenario where only one token probability is significantly large. Specifically, denote p_{max} as the largest probability of a token in the probability distribution $P_M(\cdot | x^{-N_p:(t-1)})$. We make an intuitive assumption that except p_{max} , other $|\mathcal{V}| - 1$ probabilities are small enough to uniformly fill in the remaining $1 - p_{max}$ probability value.

Previous work claims that unbiased watermarks have no impact on text quality by maintaining the same expectation (Hu et al., 2024). However, we challenge this claim in the low-entropy protocol described above. We show that in such a protocol, unbiased watermarks can still affect text quality because of the risk of unsatisfactory outputs. Consider the following example.

Example 1. Assuming that the token set only consists of two tokens $\mathcal{V} = \{A, B\}$, at a typical step, an LLM outputs the probability of generat-

ing A (p_A) and B (p_B) as $(p_A, p_B) = (0.8, 0.2)$. Consider the following two unbiased watermarks. W_1 : with a probability of 0.8 always generating A and with a probability of 0.2 always generating B ; W_2 : with a probability of 0.5, the probability distribution becomes $(p_A, p_B) = (0.9, 0.1)$ and with the other probability of 0.5, it becomes $(p_A, p_B) = (0.7, 0.3)$.

In Example 1, one can view the prompt as ‘The quick brown fox jumps over the lazy’, A as the token ‘dog’, and B as all other tokens. It is easy to show that watermarks W_1 and W_2 are both unbiased. However, risk-averse people (Pratt, 1978) will prefer watermark W_2 because W_2 does not have a possibility that only B is sampled. B represents unsatisfactory outputs in low-entropy scenarios which could significantly harm text quality, and we want the risk of sampling B to be as low as possible.² At any generation step, let x_{max} denote the token with the maximum probability p_{max} . We measure the risk by the variance (Sharpe, 1998) of $p_{max}^{w,k}$ among watermark keys, where $p_{max}^{w,k}$ denotes the altered value of p_{max} with a watermarking method and a key k . We show that STA-1 has a lower risk compared to previous unbiased watermarks in the following theorem. **To put it plainly, under the same expectation, the variance of the altered probabilities (risk) by STA-1 is lower.**

Theorem 2. *Assume $1 - \alpha \leq p_{max} < 1$, where α represents the partition hyperparameter used in Dipmark. For the low-entropy protocol above, the STA-1 method has a lower variance in the probability of generating x_{max} compared to other unbiased methods (including Dipmark, γ -reweight, and RDW) (Hu et al., 2024; Wu et al., 2024; Kuditipudi et al., 2023). Formally,*

$$\begin{aligned} \mathbb{V}_{k \sim P_K(k)}^{\text{STA-1}} \left[p_{max}^{w,k} \right] &< \mathbb{V}_{k \sim P_K(k)}^{\text{Dipmark}} \left[p_{max}^{w,k} \right] \\ &= \mathbb{V}_{k \sim P_K(k)}^{\gamma\text{-reweight}} \left[p_{max}^{w,k} \right] < \mathbb{V}_{k \sim P_K(k)}^{\text{RDW}} \left[p_{max}^{w,k} \right], \end{aligned} \quad (2)$$

for any $\alpha \in [0, 0.5]$ used in Dipmark.

Proof. See Appendix C.2.

3.1.2 Statistical Test Guarantees

The proposed STA-1 method also has a statistical test guarantee of type II error for detection. Specifically, the detection of STA-1 compares the empirical proportion of green list tokens in the given

²We refer readers to Appendix D for a conventional example in finance and a better understanding of the analysis via utility theory.

text against the green list proportion γ (Kirchenbauer et al., 2023a). We employ the z -test where the null hypothesis (H_0) is that the text is generated without knowing the green-red list partition. Denote $|S|_G$ as the number of green list tokens in this text. Under H_0 , $|S|_G$ follows a Binomial distribution $B(T, \gamma)$ with a mean of γT and a variance of $\gamma(1 - \gamma)T$. The z -score is calculated with the empirical $|S|_G$ as

$$z = \frac{|S|_G - \gamma T}{\sqrt{\gamma(1 - \gamma)T}}. \quad (3)$$

The alternative hypothesis (H_a) is that the text is generated with STA-1. Under H_a , $|S|_G$ is expected to be larger than γT . We can detect watermarked texts with a certain confidence level if the z -score exceeds a z threshold.

To ensure the effectiveness of the z -test, under H_a , a lower bound on the expectation of $|S|_G$ and an upper bound on the variance of $|S|_G$ are required. We establish the necessary lower and upper bounds in the following theorem. Because both bounds are related to the Gini index of the LLM output distribution, we define the Gini index first.

Definition 2 (Gini index). Given a discrete probability distribution $p = (p_1, p_2, \dots, p_N)$, the Gini index of p is defined as

$$Gini(p) = \sum_{i=1}^N p_i(1 - p_i). \quad (4)$$

A low Gini index implies less uncertainty in the probability distribution, resulting in a low-entropy scenario. Next, we propose the mean and variance bounds of $|S|_G$.

Theorem 3. For STA-1 generated text sequences with T tokens, let the random green list have a fixed size of $\gamma|\mathcal{V}|$, and p_i^t denote the LLM’s raw output probability of the i -th token in \mathcal{V} at step t , $i = 1, 2, \dots, |\mathcal{V}|$, $p^t = (p_1^t, p_2^t, \dots, p_{|\mathcal{V}|}^t)$. If an STA-1 generated sequence S has an average Gini index larger than or equal to $Gini^*$, that is, $\frac{1}{T} \sum_{t=1}^T Gini(p^t) \geq Gini^*$, then the expectation of $|S|_G$ is at least

$$\mathbb{E}(|S|_G) \geq \gamma T + (1 - \gamma)\gamma T Gini^*. \quad (5)$$

With one additional assumption that γ and $Gini^*$ satisfy $\gamma + (1 - \gamma)\gamma Gini^* \geq 0.5$, the variance of $|S|_G$ is at most

$$\mathbb{V}(|S|_G) \leq T[\gamma + (1 - \gamma)\gamma Gini^*][1 - \gamma - (1 - \gamma)\gamma Gini^*]. \quad (6)$$

Proof. See Appendix C.3.

Remark 1. The additional assumption required for the variance upper bound, $\gamma + (1 - \gamma)\gamma Gini^* \geq 0.5$, implies that a larger green list is necessary in low-entropy scenarios to establish an upper bound on the variance of $|S|_G$. By selecting $\gamma \geq 0.5$, this assumption holds for any $Gini^*$.

Remark 2. Compared to the Spike entropy proposed by Kirchenbauer et al. (2023a), the Gini index is a commonly used metric in machine learning to measure the uncertainty of a probability distribution, such as CART decision tree (Breiman, 2017).

Having established the mean and variance bounds for $|S|_G$, with an additional condition, we derive from Theorem 3 a corollary that provides an explicit upper bound on the type II error of the z -test in detecting STA-1.

Corollary 1. Given that Theorem 3 holds, if $Gini^* > \tilde{z}/\sqrt{\gamma(1 - \gamma)T}$, we have the type II error $\beta = P(\frac{|S|_G - \gamma T}{\sqrt{\gamma(1 - \gamma)T}} \leq \tilde{z} | H_a)$ satisfy

$$\beta \leq \frac{\bar{\mathbb{V}}}{\bar{\mathbb{V}} + (\underline{\mathbb{E}} - \gamma T - \tilde{z}\sqrt{\gamma(1 - \gamma)T})^2}, \quad (7)$$

where \tilde{z} is the z threshold value, $\underline{\mathbb{E}}$ and $\bar{\mathbb{V}}$ are the lower bound and upper bound values on $\mathbb{E}(|S|_G)$ and $\mathbb{V}(|S|_G)$ as established in Theorem 3, respectively.

Proof. See Appendix C.4.

A higher $Gini^*$ increases $\underline{\mathbb{E}}$ and decreases $\bar{\mathbb{V}}$, resulting in a reduced upper bound on the type II error. Therefore, the test has higher statistical power in high-entropy scenarios.

3.2 Sampling M Then Accepting (STA-M)

A low-entropy scenario indicates a low Gini index which weakens the watermark strength based on Theorem 3. To enhance the watermark strength, we propose the Sampling M Then Accepting (STA-M) method, an extension of STA-1. STA-M employs a heuristic threshold τ for entropy at each generation step. In detail, at generation step t , we first calculate the entropy τ^t of the probability distribution $P_M(\cdot | x^{-N_p:(t-1)})$. If it shows low entropy $\tau^t \leq \tau$, we apply STA-1 at this generation step; if it shows high entropy $\tau^t > \tau$, we repeat sampling if the previously sampled token is in the red list, and the procedure repeats at most M times.

The detailed algorithm and analysis of STA-M can be found in Appendix E. According to Remark 3 in Appendix E, STA-M is biased. In low-entropy

steps where probabilities are concentrated on a few tokens, actively using STA-M by repeated sampling can skew these probabilities, thereby reducing text quality. On the contrary, in high-entropy steps, since there are more acceptable tokens, the impact of repeated sampling on text quality is weakened. Therefore, STA-M only repeats sampling in high-entropy steps, which could increase watermark strength and largely maintain text quality.

4 Experiments

In this section, we conducted computational experiments to evaluate the performance of STA-1 and STA-M using two public datasets. We benchmarked our methods against various watermarking baselines on text quality, watermark strength, and detection time. Moreover, we discussed the risk of unsatisfactory outputs in the low-entropy dataset. Finally, we conducted a robustness analysis of STA-1 and STA-M against different watermarking attacks.

4.1 Experimental Setup

Datasets and metrics. We employed two public datasets: C4 subset (Raffel et al., 2020) for news-like (high-entropy) text generation and HumanEval (Chen et al., 2021) for code (low-entropy) generation. We evaluated the performance of different watermarking methods on text quality and watermark strength. For text quality, we measured perplexity (PPL) and coherence (Gao et al., 2021) for generations on C4; We computed PPL and pass@ k scores of code generations (Chen et al., 2021) for HumanEval. We refer readers to Appendix F.1 for more dataset details and prompts used in each dataset. For watermark strength, we set z thresholds as 2 and 2.5 and report the F1-score and AUC of watermark detection. Additionally, for the C4 subset, we employed true positive rate at false positive rate (TPR@FPR) as another metric to evaluate the detection (Liu et al., 2023a).

Baselines. We chose KGW (Kirchenbauer et al., 2023a), SWEET (Lee et al., 2023), and EWD (Lu et al., 2024a) as the biased watermark baselines. Additionally, we selected RDW (Kuditipudi et al., 2023), Dipmark (Wu et al., 2024), and γ -reweight (Hu et al., 2024) as the unbiased watermark baselines. Specifically, we set KGW, SWEET, and EWD with a fixed green list proportion $\gamma = 0.5$. For KGW, we employed diverse logit increments $\delta \in \{1, 1.5, 2\}$. For SWEET, we fixed the logit

increment as $\delta = 2$ (Lee et al., 2023). The entropy threshold of SWEET was set the same as STA-M for a fair comparison. For EWD, the spike entropy parameter was set according to their public implementation.³ We set the watermark key length as 256 in RDW. The partition parameter of Dipmark was set as $\alpha \in \{0.3, 0.4, 0.5\}$. When $\alpha = 0.5$, we report this result as γ -reweight. Note that γ -reweight (Hu et al., 2024) does not include a z -test. Therefore, we implemented the z -score in Dipmark (Wu et al., 2024) for γ -reweight by counting the number of tokens in the latter portion of the token set. Also, RDW only contains a permutation test that reports p-values. We set p-value thresholds at 0.05 and 0.01 to approximate two z -tests.

Implementation details. We utilized different variants of LLaMA-2-7B (Touvron et al., 2023) as our generative LLMs, and LLaMA-2-13B to compute perplexity. For hyperparameters in STA-M, we set $M \in \{4, 8, 16\}$ and two entropy thresholds τ for different datasets. We conducted a robustness check on τ in Appendix F.2 and selected different τ s for different datasets in the final experiment. For each method, we run 10 times to conduct all pairwise statistical tests. Results in the following tables show only average values. For detection efficiency, we report the detection time for all generations. We refer readers to Appendix F.1 for more details on implementation.

4.2 Results on C4

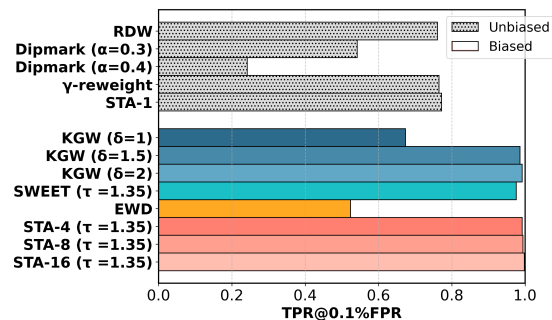


Figure 1: Result Comparison of Watermark Strength of TPR@0.1%FPR Between Our Method and Baselines for the C4 Dataset.

For the C4 dataset, each method generates at least 500 text sequences with at least 200 ± 5 tokens (Kirchenbauer et al., 2023a). Table 2 demonstrates each method’s text quality, watermark strength, and detection efficiency for 500 generations, and we present generated text examples

³<https://github.com/luyijian3/EWD>

Table 2: Result Comparison between Our Methods and Baselines on Text Quality and Watermark Strength for the C4 Dataset. For unbiased watermarks, the best results without statistical differences are underlined. For biased watermarks, the best results without statistical differences are shown in **bold**.

	Method	Text Quality		Watermark Strength				Detection Efficiency
		\downarrow PPL	\uparrow Coherence	$z = 2.0$		$z = 2.5$		Total Time
				\uparrow F1	\uparrow AUC	\uparrow F1	\uparrow AUC	
	No Watermark	7.474	0.604	0.046	0.500	0.012	0.500	46s
Unbiased	RDW	7.650	0.592	0.942	0.942	0.948	0.950	4h
	Dipmark($\alpha=0.3$)	<u>7.415</u>	<u>0.599</u>	0.933	0.935	0.909	0.915	44s
	Dipmark($\alpha=0.4$)	7.384	<u>0.601</u>	<u>0.957</u>	<u>0.957</u>	0.954	0.955	44s
	γ -reweight	<u>7.436</u>	<u>0.599</u>	<u>0.961</u>	<u>0.961</u>	<u>0.963</u>	<u>0.963</u>	44s
	STA-1	<u>7.387</u>	<u>0.600</u>	<u>0.962</u>	<u>0.961</u>	<u>0.963</u>	<u>0.963</u>	46s
Biased	KGW($\delta=1$)	7.591	0.601	0.961	0.962	0.940	0.944	46s
	KGW($\delta=1.5$)	7.844	0.600	0.985	0.984	0.990	0.990	46s
	KGW($\delta=2$)	8.091	0.595	0.986	0.986	0.992	0.992	46s
	SWEET($\tau=1.35$)	7.917	0.600	0.980	0.980	0.989	0.989	46s
	EWD	7.580	0.606	0.930	0.932	0.880	0.892	46s
	STA-4($\tau=1.35$)	7.611	0.599	0.973	0.972	0.988	0.988	46s
	STA-8($\tau=1.35$)	8.006	0.592	0.975	0.975	0.987	0.987	46s
	STA-16($\tau=1.35$)	8.199	0.588	0.973	0.972	0.988	0.988	46s

in Appendix F.3. We can observe that the proposed STA-1 method achieves comparable perplexity and coherence when compared to no watermark generation and existing unbiased watermarks, including RDW, Dipmark, and γ -reweight. **This result empirically shows that the STA-1 method is unbiased.** In terms of watermark strength, the STA-1 method also achieves satisfactory results on F1 and AUC and is not inferior to existing unbiased benchmarks. We also plot the detection performance of TPR@0.1%FPR in Figure 1. We observe that, unlike some unbiased watermarks such as Dipmark, which experience a significant drop in TPR@0.1%FPR, the STA-1 method remains comparable to the best-performing unbiased watermark RDW. Furthermore, based on Table 2, **the STA-1 method is highly efficient**, taking only 46 seconds to detect 500 generations, while RDW requires 4 hours to detect the same number of generations.

In Table 2, we also report the performance of STA-M, which samples more times at high-entropy steps for improving watermark strength. In terms of watermark strength, it is evident that STA-M ($M \in \{4, 8, 16\}$) outperforms all unbiased watermarks and demonstrates results comparable to the biased KGW watermark ($\delta \in \{1.5, 2\}$) and SWEET. We also plot the TPR@0.1%FPR scores for different parameter settings of STA-M and biased watermark baselines in Figure 1. From this, we observe that the watermark strength of EWD and SWEET is inferior to our method. Meanwhile,

the watermark strength of KGW ($\delta \in \{1, 1.5, 2\}$) varies significantly, ranging from 0.7 to 0.99. In contrast, our STA-M ($M \in \{4, 8, 16\}$) method remains stable across all parameter settings, with a consistent TPR@0.1%FPR score over 0.99. Regarding text quality, STA-M does not experience significant drops compared to the unbiased watermarks and remains comparable to the biased KGW methods.

4.3 Results on HumanEval

We then compare our methods against baselines on the HumanEval dataset, a low-entropy code generation benchmark. We report perplexity, pass@k scores, and watermark strength for all methods in Table 3. Since it is preferable not to control the length of code during generation, we remove detection time results. We observe that our STA-1 method achieves similar perplexity, pass@k scores, and watermark strength compared to other unbiased watermarking methods. This also empirically corroborates that the STA-1 method is unbiased.

Moreover, we examine the risk of unsatisfactory outputs produced by unbiased watermarks for low-entropy generations. Specifically, we ran 10 times of code generation for each problem using different unbiased watermarking methods with 10 different keys. We compute the average variance of perplexity for each problem, as well as the average number of passed codes among all passed problems. The results are shown in Figure 2. From the

Table 3: Result Comparison between Our Methods and Baselines on Text Quality and Watermark Strength for the HumanEval Dataset. For unbiased watermarks, the best results without statistical differences are underlined. For biased watermarks, the best results without statistical differences are shown in **bold**.

	Method	Text Quality				Watermark Strength			
		\downarrow PPL	\uparrow Pass@1	\uparrow Pass@5	\uparrow Pass@10	$z = 2.0$	$z = 2.5$	\uparrow F1	\uparrow AUC
	No Watermark	3.041	0.138	0.405	0.537	0.114	0.494	0.072	0.497
Unbiased	RDW	<u>3.159</u>	0.134	0.362	0.470	0.408	0.628	0.343	0.604
	Dipmark($\alpha=0.3$)	<u>3.037</u>	0.144	<u>0.392</u>	<u>0.512</u>	0.518	0.665	0.423	0.625
	Dipmark($\alpha=0.4$)	<u>3.101</u>	0.141	<u>0.393</u>	<u>0.512</u>	0.516	0.668	0.429	0.634
	γ -reweight	<u>3.088</u>	<u>0.142</u>	0.371	0.488	<u>0.522</u>	<u>0.671</u>	<u>0.479</u>	<u>0.655</u>
	STA-1	<u>3.006</u>	<u>0.147</u>	<u>0.394</u>	<u>0.494</u>	<u>0.526</u>	<u>0.677</u>	<u>0.472</u>	<u>0.651</u>
Biased	KGW($\delta=1$)	3.078	0.135	0.326	0.415	0.471	0.643	0.416	0.627
	KGW($\delta=1.5$)	3.499	0.098	0.308	0.427	0.720	0.770	0.650	0.730
	KGW($\delta=2$)	3.723	0.098	0.254	0.372	0.737	0.775	0.733	0.785
	SWEET($\tau=1.95$)	3.125	0.127	0.312	0.402	0.386	0.605	0.299	0.583
	EWD	3.106	0.132	0.335	0.439	0.469	0.630	0.385	0.607
	STA-4($\tau=1.95$)	3.175	0.135	0.392	0.500	0.633	0.685	0.594	0.679
	STA-8($\tau=1.95$)	2.842	0.146	0.399	0.537	0.652	0.703	0.587	0.675
	STA-16($\tau=1.95$)	3.024	0.140	0.382	0.476	0.725	0.764	0.640	0.717

figure, it is evident that the STA-1 method demonstrates the lowest variance of perplexity compared to RDW, Dipmark($\alpha=0.3$), Dipmark($\alpha=0.4$), and γ -reweight. A lower variance indicates a lower risk among different text generations under different keys. Additionally, we observe that for STA-1, the average number of passed codes among all passed problems is significantly larger than that of other unbiased watermarks, exceeding 3.1, while others remain below 2.9. Therefore, we can conclude that **the STA-1 method has a lower risk when generating low-entropy texts** compared to existing unbiased watermarks, as discussed in Theorem 2.

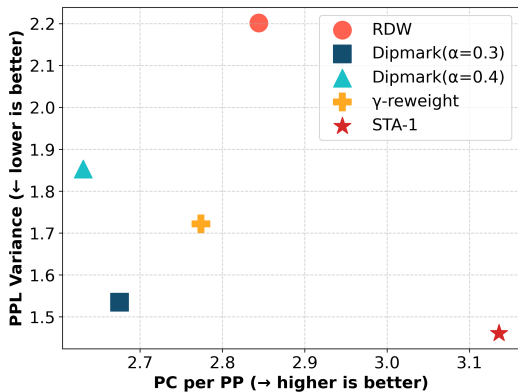


Figure 2: Comparison on the Risk of Unsatisfactory Outputs for Unbiased Watermarks. For space concerns, we denote the average number of passed codes among all passed problems as PC per PP.

We then report the performance of STA-M in

Table 3. As shown, by repeating sampling during high-entropy steps, STA-M ($M \in 4, 8, 16$) achieves higher watermark strength compared to all unbiased watermarks, while maintaining similar pass scores. Specifically, the STA-16 method achieves comparable watermark strength against biased watermark KGW($\delta = 2$) with an AUC of 0.764 ($z = 2$) against 0.775. Meanwhile, the text quality is maintained with a pass@10 of 0.476, highlighting the efficacy of the heuristics to enhance watermark strength at high-entropy generation steps.

4.4 Attacking STA

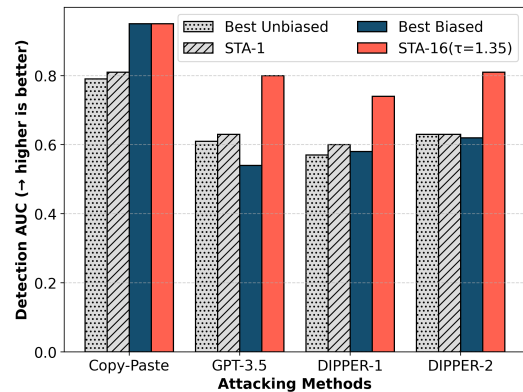


Figure 3: Attacking Watermarks for C4. For baselines, we report the *highest* AUC score of unbiased and biased watermarks against each attack. Full results and discussions are available in Appendix F.4 and Table 6.

We assess the robustness of different watermarking methods under various attacks, including the copy-paste attack (Kirchenbauer et al., 2023a), paraphrasing using GPT-3.5, and two configurations of the DIPPER attack (Krishna et al., 2024). For space concerns, we describe the detailed attack setting and report the F1-score and AUC of watermark detection with $z = 2$ in Appendix F.4. We plot the AUC of watermark detection with $z = 2$ for STA-1 and STA-16, alongside the *highest* AUC values of biased and unbiased benchmarks against each attack in Figure 3. As depicted, on the one hand, the unbiased STA-1 method achieves satisfactory performance, matching the best-performing unbiased benchmark in each attack. This empirically demonstrates that **the STA-1 is also robust to various attacks**. On the other hand, by repeating sampling as high-entropy steps, the STA-16 method achieves better robustness than KGW.

We detail the reasons for the robustness of STA-1 and STA-M as follows. For the copy-paste attack, since our method inherits from KGW, it is naturally robust to simple text insertion and removal (Kirchenbauer et al., 2023a). Meanwhile, LLM-based attacks, such as GPT-3.5 and DIPPER, are designed to replace tokens in given texts by sampling from the LLM. STA-M effectively increases the proportion of green-list tokens by raising their probability in high-entropy scenarios without compromising too much text quality, making it difficult for LLM-based attacks to replace a substantial number of tokens in STA-M-generated text and remove the watermark.

5 Related Work

With the development of LLMs, the idea of watermarking LLMs has been proposed (Aaronson, 2022) and widely explored (Tu et al., 2024). Existing white-box watermarking techniques can be categorized into watermarking during logits and probabilities generation (Wang et al., 2023; Zhao et al., 2023; Yoo et al., 2023; Ren et al., 2023; Takezawa et al., 2023; Lu et al., 2024b), and watermarking by controlling sampling strategies (Christ et al., 2023; Kuditipudi et al., 2023; Hou et al., 2023; Fairoze et al., 2023). We refer readers to Appendix G for a detailed related work.

6 Conclusions

In this work, we propose a novel watermarking method named STA-1. Theoretically, we show that

STA-1 is unbiased and has a lower risk than existing unbiased watermarks. During detection, STA-1 also provides statistical test guarantees on the type II error of watermark detection, demonstrates high efficiency in detection time, and remains robust against various watermarking attacks. Experimental results on public datasets show that STA-1 achieves the above properties simultaneously. We also extend STA-1 to STA-M, which can enhance watermark strength with small text quality shifts.

7 Limitations

We acknowledge several limitations in this work and suggest directions accordingly for future improvement. First, watermarking low-entropy tasks remains challenging, and future work could devise better watermarking methods to improve watermark strength while maintaining text quality. Second, future work could incorporate more datasets and models to evaluate our method. Third, LLM watermarks should be robust against paraphrasing attacks like GPT-3.5 and DIPPER even in low-entropy scenarios. Future work can consider extending watermarking methods by enhancing robustness in these scenarios. Fourth, it may also be useful to consider context code history (Hu et al., 2024) to extend the unbiased results from the token level to the sequence level.

8 Ethical Statement

Watermarking methods for LLMs are designed to enhance accountability in their deployment by facilitating the precise and efficient detection of LLM-generated content. While showcasing broad benefits, these watermarking techniques also carry inherent risks: if the underlying watermarking mechanism (which is typically kept confidential by the LLM owner) is exposed, malicious actors could exploit it to manipulate watermarks, such as falsely attributing LLM-generated content or escaping detection entirely. Therefore, to maintain ethical standards, users must protect the confidentiality of the watermarking mechanism (such as the hash function and the key required in our proposed method) and acknowledge that watermarks alone cannot prevent all types of misuse. Moreover, a watermarking method should be implemented alongside broader safeguards to minimize unintended harms, such as access controls and misuse monitoring, ensuring it functions as a tool for accountability rather than a means for new abuses.

References

- Scott Aaronson. 2022. My ai safety lecture for ut effective altruism. <https://scottaaronson.blog/?p=6823>. Accessed: 2024-05-15.
- Leo Breiman. 2017. *Classification and regression trees*. Routledge.
- Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. 2021. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*.
- Miranda Christ, Sam Gunn, and Or Zamir. 2023. Undetectable watermarks for language models. *arXiv preprint arXiv:2306.09194*.
- Sumanth Dathathri, Abigail See, Sumedh Ghaisas, Po-Sen Huang, Rob McAdam, Johannes Welbl, Vandana Bachani, Alex Kaskasoli, Robert Stanforth, Tatiana Matejovicova, et al. 2024. Scalable watermarking for identifying large language model outputs. *Nature*, 634(8035):818–823.
- Gerard Debreu et al. 1954. Representation of a preference ordering by a numerical function. *Decision processes*, 3:159–165.
- Jaiden Fairoze, Sanjam Garg, Somesh Jha, Saeed Mahlouljifar, Mohammad Mahmood, and Mingyuan Wang. 2023. Publicly detectable watermarking for language models. *arXiv preprint arXiv:2310.18491*.
- Xiao Fang, Shangkun Che, Minjia Mao, Hongzhe Zhang, Ming Zhao, and Xiaohang Zhao. 2024. Bias of ai-generated content: an examination of news produced by large language models. *Scientific Reports*, 14(1):5224.
- Pierre Fernandez, Antoine Chaffin, Karim Tit, Vivien Chappelier, and Teddy Furon. 2023. Three bricks to consolidate watermarks for large language models. In *2023 IEEE International Workshop on Information Forensics and Security (WIFS)*, pages 1–6. IEEE.
- Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. Simcse: Simple contrastive learning of sentence embeddings. *arXiv preprint arXiv:2104.08821*.
- Abe Bohan Hou, Jingyu Zhang, Tianxing He, Yichen Wang, Yung-Sung Chuang, Hongwei Wang, Lingfeng Shen, Benjamin Van Durme, Daniel Khashabi, and Yulia Tsvetkov. 2023. Semstamp: A semantic watermark with paraphrastic robustness for text generation. *arXiv preprint arXiv:2310.03991*.
- Zhengmian Hu, Lichang Chen, Xidong Wu, Yihan Wu, Hongyang Zhang, and Heng Huang. 2024. Unbiased watermark for large language models. In *The Twelfth International Conference on Learning Representations, ICLR 2024*. OpenReview.net.
- John Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, and Tom Goldstein. 2023a. A watermark for large language models. In *International Conference on Machine Learning*, 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. 2024. Paraphrasing evades detectors of ai-generated text, but retrieval is an effective defense. *Advances in Neural Information Processing Systems*, 36.
- Rohith Kuditipudi, John Thickstun, Tatsunori Hashimoto, and Percy Liang. 2023. Robust distortion-free watermarks for language models. *arXiv preprint arXiv:2307.15593*.
- Taehyun Lee, Seokhee Hong, Jaewoo Ahn, Ilgee Hong, Hwaran Lee, Sangdoon Yun, Jamin Shin, and Gunhee Kim. 2023. Who wrote this code? watermarking for code generation. *arXiv preprint arXiv:2305.15060*.
- Aiwei Liu, Leyi Pan, Xuming Hu, Shiao Meng, and Lijie Wen. 2023a. A semantic invariant robust watermark for large language models. In *The Twelfth International Conference on Learning Representations*.
- Aiwei Liu, Leyi Pan, Yijian Lu, Jingjing Li, Xuming Hu, Lijie Wen, Irwin King, and Philip S Yu. 2023b. A survey of text watermarking in the era of large language models. *arXiv preprint arXiv:2312.07913*.
- Yijian Lu, Aiwei Liu, Dianzhi Yu, Jingjing Li, and Irwin King. 2024a. An entropy-based text watermarking detection method. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11724–11735.
- Yijian Lu, Aiwei Liu, Dianzhi Yu, Jingjing Li, and Irwin King. 2024b. An entropy-based text watermarking detection method. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 11724–11735. Association for Computational Linguistics.
- Yisroel Mirsky, Ambra Demontis, Jaidip Kotak, Ram Shankar, Deng Gelei, Liu Yang, Xiangyu Zhang, Maura Pintor, Wenke Lee, Yuval Elovici, et al. 2023. The threat of offensive ai to organizations. *Computers & Security*, 124:103006.
- Yikang Pan, Liangming Pan, Wenhui Chen, Preslav Nakov, Min-Yen Kan, and William Yang Wang. 2023. On the risk of misinformation pollution with large language models. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

- John W Pratt. 1978. Risk aversion in the small and in the large. In *Uncertainty in economics*, pages 59–79. Elsevier.
- Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.
- 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. *Journal of machine learning research*, 21(140):1–67.
- Jie Ren, Han Xu, Yiding Liu, Yingqian Cui, Shuaiqiang Wang, Dawei Yin, and Jiliang Tang. 2023. A robust semantics-based watermark for large language model against paraphrasing. *arXiv preprint arXiv:2311.08721*.
- Baptiste Roziere, Jonas Gehring, Fabian Gloeckle, Sten Sootla, Itai Gat, Xiaoqing Ellen Tan, Yossi Adi, Jingyu Liu, Tal Remez, Jérémy Rapin, et al. 2023. Code llama: Open foundation models for code. *arXiv preprint arXiv:2308.12950*.
- William F Sharpe. 1998. The sharpe ratio. *Streetwise—the Best of the Journal of Portfolio Management*, 3:169–185.
- Yuki Takezawa, Ryoma Sato, Han Bao, Kenta Niwa, and Makoto Yamada. 2023. Necessary and sufficient watermark for large language models. *arXiv preprint arXiv:2310.00833*.
- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.
- Shangqing Tu, Yuliang Sun, Yushi Bai, Jifan Yu, Lei Hou, and Juanzi Li. 2024. [Waterbench: Towards holistic evaluation of watermarks for large language models](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, ACL 2024, Bangkok, Thailand, August 11-16, 2024, pages 1517–1542. Association for Computational Linguistics.
- Lean Wang, Wenkai Yang, Deli Chen, Hao Zhou, Yankai Lin, Fandong Meng, Jie Zhou, and Xu Sun. 2023. Towards codable text watermarking for large language models. *arXiv preprint arXiv:2307.15992*.
- Wikipedia. 2024. St. Petersburg paradox — Wikipedia, the free encyclopedia. <http://en.wikipedia.org/w/index.php?title=St.%20Petersburg%20paradox&oldid=1212997265>. [Online; accessed 21-May-2024].
- Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2019. Huggingface’s transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.
- Yihan Wu, Zhengmian Hu, Junfeng Guo, Hongyang Zhang, and Heng Huang. 2024. A resilient and accessible distribution-preserving watermark for large language models. In *International Conference on Machine Learning*.
- KiYoon Yoo, Wonhyuk Ahn, and Nojun Kwak. 2023. Advancing beyond identification: Multi-bit watermark for language models. *arXiv preprint arXiv:2308.00221*.
- Xuandong Zhao, Yu-Xiang Wang, and Lei Li. 2023. Protecting language generation models via invisible watermarking. In *International Conference on Machine Learning*, pages 42187–42199. PMLR.

A Research Gap Summary

Existing unbiased watermarks can be categorized according to the stage where watermarks are injected: distribution reweighting and controlled sampling (Liu et al., 2023b). For distribution reweighting, Hu et al. (2024) proposes γ -reweight, which uses the log-likelihood ratio (LLR) test by comparing the likelihood of the text produced by watermarked and unwatermarked white-box LLMs. It requires the prompt as input and a white-box LLM in watermark detection (Fernandez et al., 2023; Hu et al., 2024). Also, the watermark is unstable because changing the first token of the generated text can lead to huge deviations from the original likelihood value (Fernandez et al., 2023). In response, Wu et al. (2024) avoid the LLR test and propose Dipmark, an extension of γ -reweight with more general parameter settings. However, although both γ -reweight and Dipmark ensure the type I error of watermark detection, they fail to provide statistical guarantees for the type II error (Hu et al., 2024; Wu et al., 2024). For controlled sampling, Christ et al. (2023) introduce a watermarking method that uses a sequence of random values to guide the token sampling process. However, their method is not robust enough against simple removal attacks (Liu et al., 2023b). Kuditipudi et al. (2023) also use random values to control the sampling and introduce a permutation test on detection that does not require white-box access

to LLMs. However, this permutation test is time-consuming theoretically and empirically. Dathathri et al. (2024) propose SynthID-Text, which is an unbiased watermarking method and incorporates a tournament-based sampling technique. With a tournament hyperparameter m , 2^m tokens are sampled and split into pairs. m rounds of tournaments are completed by different random watermarking functions with different random seeds. It is not cost-efficient for local usage and is not robust against strong paraphrasing attacks. Fairoze et al. (2023) propose to sample the token sequence generation until its hash matches a key value. According to their distortion-free definition, the upper bound of the difference between probabilities before and after watermarking is $\exp(-a)$, where a is the minimal entropy. The difference is not negligible in low-entropy scenarios. Note that using random values to control sampling can be treated as a special case of distribution reweighting where only the probability of the sampled token is reweighted to 1 (Kuditipudi et al., 2023). Thus, we build our analysis framework in Section 2 solely based on distribution reweighting.

B Details of Previous Methods

Distribution reweighting refers to methods that adjust the output distribution $P_M(x^t|x^{-N_p:(t-1)})$ at each step t by artificially increasing probabilities for certain tokens while reducing those for others. The direction and magnitude (increasing or decreasing) of change in probability mass for a token are determined by the private key k .

KGW (Kirchenbauer et al., 2023a) first randomly splits the vocabulary set \mathcal{V} into two non-overlapping lists based on a uniformly distributed key k : a ‘green’ list and a ‘red’ list. This method has two versions: the ‘hard’ version completely ignores the red list tokens and only samples tokens from the green list; The ‘soft’ version adds a predefined constant δ to logits of green list tokens while keeping logits of red list tokens fixed. The soft KGW reweights distribution as

$$P_{M,w}(x^t = j|x^{-N_p:(t-1)}; k) = \frac{\exp(l_j^t + \mathbb{1}_{\text{Green}}(j)\delta)}{\sum_{i \in \text{Red}} \exp(l_i^t) + \sum_{i \in \text{Green}} \exp(l_i^t + \delta)},$$

where j denotes the j -th token within the vocabulary set, l_j^t is its logit output by the original LLM at step t , and $\mathbb{1}_{\text{Green}}(j)$ is an indicator function hav-

ing a value of 1 when j is in the green list and 0 otherwise.

Wu et al. (2024) propose an unbiased reweighting method, named Dipmark. Dipmark arranges all probability masses over the vocabulary set from the original LLM output consecutively within the interval $[0, 1]$ and then randomly permutes their orders based on a key k . A hyperparameter $\alpha \in [0, 0.5]$ partitions the probability interval $[0, 1]$ into three segments: $[0, \alpha]$, $(\alpha, 1 - \alpha]$, and $(1 - \alpha, 1]$. Probability masses in the first segment are set to 0, those in the second remain constant, and those in the third are doubled. Denote the token order after permutation as $\tilde{\mathcal{V}}$, the adjusted probability for the j -th token within $\tilde{\mathcal{V}}$ is $P_{M,w}(x^t = j|x^{-N_p:(t-1)}; k) = F(j|\tilde{\mathcal{V}}) - F(j-1|\tilde{\mathcal{V}})$, with $F(j|\tilde{\mathcal{V}})$ being defined as

$$F(j|\tilde{\mathcal{V}}) = \max \left[\sum_{i \in \tilde{\mathcal{V}}: i \leq j} P_M(x^t = i|\cdot) - \alpha, 0 \right] + \max \left[\sum_{i \in \tilde{\mathcal{V}}: i \leq j} P_M(x^t = i|\cdot) - (1 - \alpha), 0 \right].$$

Notably, Dipmark becomes γ -reweight (Hu et al., 2024) when $\alpha = 0.5$.

Another unbiased reweighting method, RDW (robust distortion-free watermark), is developed by Kuditipudi et al. (2023). We focus on the RDW method with an inverse transform sampling scheme. In RDW, the uniformly random key $k = (\Pi, u)$, where Π represents a random shuffle of all probability masses $P_M(x^t|x^{-N_p:(t-1)})$ over the vocabulary set within the interval $[0, 1]$, and u is a random value following the distribution $U(0, 1)$. RDW first permutes the order of all $P_M(x^t|x^{-N_p:(t-1)})$ within the interval $[0, 1]$ according to Π , then it utilizes u as the cumulative distribution function value of $P_M(x^t|x^{-N_p:(t-1)})$ with respect to the permutation. Let $\Pi(j)$ denote the j -th token in the ordered vocabulary set under the permutation Π . Following the inverse transform sampling scheme, the value u is inverse transformed to generate a token through

$$x^s = \Pi(\min\{j : \sum_{i=1}^j P_M(x^t = \Pi(i)|x^{-N_p:(t-1)}) \geq u\}),$$

where x^s is the sampled token. Therefore, we have $P_{M,w}(x^t = x^s|x^{-N_p:(t-1)}; k) = 1$, and the probabilities of all other tokens are reweighted to 0 accordingly.

C Proofs

C.1 Proof of Theorem 1

To simplify notation, we denote the size of the vocabulary set $|\mathcal{V}|$ as N , the size of the green list as N_G , and the size of the red list as N_R . Given the proportion of green list γ , we have $N_G = \gamma N$ and $N_R = (1 - \gamma)N$. At a generation step, let $p = (p_1, p_2, \dots, p_N)$ denote the raw probability output by the LLM over the vocabulary set. Let j represent a token within the vocabulary set, $j \in (1, 2, \dots, N)$. We denote by $p_j^{w,k}$ the adjusted probability of token j under the STA-1 watermarking method with key k . The key k is sampled randomly from a uniform distribution $P_K(k)$.

To conveniently compute $\mathbb{E}_{k \sim P_K(k)} [p_j^{w,k}]$, we consider the uniformly random partition of green and red lists associated with the uniformly distributed key k as the following process. Initially, token j is randomly assigned to the green list with a probability of γ and to the red list with a probability of $1 - \gamma$. Subsequently, tokens are randomly sampled from the remaining pool to fill the green list, with all remaining tokens then placed in the red list. For the adjusted probability, we have

$$p_j^{w,k} = \begin{cases} p_j + (\sum_{i \in R} p_i) p_j & j \in G \\ (\sum_{i \in R} p_i) p_j & j \in R \end{cases}$$

Next, we first analyze the scenario where $j \in G$ and compute $\mathbb{E}_{G,R:j \in G} [p_j^{w,k}]$. The expectation is taken over uniformly random partitions of green/red lists that fulfill $j \in G$. Let

$$\begin{aligned} h_j(p) &= \mathbb{E}_{G,R:j \in G} [p_j^{w,k}] \\ &= \mathbb{E}_{G,R:j \in G} \left[p_j + \left(\sum_{i \in R} p_i \right) p_j \right]. \end{aligned}$$

Note that $h_j(p)$'s value remains unchanged under permutations in the order of the remaining tokens $\{p_i, i \neq j\}$. Thus, we have the equality that $h_j(p) = \mathbb{E}_{\Pi} [h_j(\Pi p_{-j})]$, where Π represents a random permutation of the remaining tokens p_{-j} while preserving the position of p_j . Since $h_j(\Pi p_{-j})$ is a linear function of p_{-j} , we then get

$$h_j(p) = \mathbb{E}_{\Pi} [h_j(\Pi p_{-j})] = h_j(\mathbb{E}_{\Pi} [\Pi p_{-j}]).$$

The expectation of the probability values at the remaining $(N - 1)$ positions over permutations of their corresponding tokens $\mathbb{E}_{\Pi} [\Pi p_{-j}]$ yields a

probability distribution \bar{p} where $\bar{p}_j = p_j$ and $\bar{p}_i = (1 - p_j)/(N - 1)$ for $i \neq j$. With this \bar{p} , we derive that

$$\begin{aligned} h_j(p) &= h_j(\bar{p}) = \mathbb{E}_{G,R:j \in G} \left[\bar{p}_j + \left(\sum_{i \in R} \bar{p}_i \right) \bar{p}_j \right] \\ &= p_j + \frac{N_R}{N - 1} (1 - p_j) p_j. \end{aligned}$$

Then, we analyze the scenario where $j \in R$ and compute $\mathbb{E}_{G,R:j \in R} [p_j^{w,k}]$. Let

$$\begin{aligned} f_j(p) &= \mathbb{E}_{G,R:j \in R} [p_j^{w,k}] \\ &= \mathbb{E}_{G,R:j \in R} \left[\left(\sum_{i \in R} p_i \right) p_j \right]. \end{aligned}$$

For the same reasons as illustrated above and using the same definition of \bar{p} , we have

$$\begin{aligned} f_j(p) &= f_j(\bar{p}) = \mathbb{E}_{G,R:j \in R} \left[\left(\sum_{i \in R} \bar{p}_i \right) \bar{p}_j \right] \\ &= \left(p_j + \frac{(N_R - 1)(1 - p_j)}{(N - 1)} \right) p_j \\ &= p_j^2 + \frac{(N_R - 1)}{N - 1} (1 - p_j) p_j. \end{aligned}$$

Finally, combining the random partition process of green and red lists described at the beginning of the proof with the derived expressions for $h_j(p)$ and $f_j(p)$, we obtain that

$$\begin{aligned} \mathbb{E}_{k \sim P_K(k)} [p_j^{w,k}] &= \gamma h_j(p) + (1 - \gamma) f_j(p) \\ &= \gamma p_j + \gamma \frac{N_R}{N - 1} (1 - p_j) p_j \\ &\quad + (1 - \gamma) p_j^2 + (1 - \gamma) \frac{(N_R - 1)}{N - 1} (1 - p_j) p_j \\ &= \left(\gamma + \frac{N_R - (1 - \gamma)}{N - 1} \right) p_j \\ &\quad + \left((1 - \gamma) - \frac{N_R - (1 - \gamma)}{N - 1} \right) p_j^2 \\ &= p_j, \end{aligned}$$

with $N_R = (1 - \gamma)N$. This concludes the proof.

C.2 Proof of Theorem 2

In this proof, we continue utilizing the notations introduced in the proof of Theorem 1 in Section C.1.

We start with the variance for the STA-1 method. Because STA-1 is an unbiased watermark by Theorem 1, we have $\mathbb{V}_{k \sim P_K(k)}^{\text{STA-1}} [p_{max}^{w,k}] =$

$\mathbb{E}_{k \sim P_K(k)}^{\text{STA-1}} \left[(p_{max}^{w,k} - p_{max})^2 \right]$. Considering the identical uniformly random partition process of green and red lists associated with the uniformly distributed key k as in the proof of Theorem 1, depending on whether the token x_{max} is assigned to the green list or not initially, $p_{max}^{w,k}$ have two possible realizations:

$$p_{max}^{w,k} = \begin{cases} p_{max} + (\sum_{i \in R} p_i) p_{max} & x_{max} \in G \\ (\sum_{i \in R} p_i) p_{max} & x_{max} \in R \end{cases}$$

Under the assumption that the probabilities of the other $N - 1$ tokens uniformly fill in the remaining $(1 - p_{max})$ probability mass, each p_i , $i \in (1, 2, \dots, N)$ and $i \neq x_{max}$, equals $(1 - p_{max})/(N - 1)$. Therefore, if $x_{max} \in G$, $p_{max}^{w,k} = p_{max} + N_R(1 - p_{max})p_{max}/(N - 1)$, and this value is fixed for all partitions of green/red lists that fulfill $x_{max} \in G$. Then we have

$$\begin{aligned} \mathbb{E}_{G,R:x_{max} \in G}^{\text{STA-1}} \left[(p_{max}^{w,k} - p_{max})^2 \right] \\ = \left[\frac{N_R(1 - p_{max})p_{max}}{(N - 1)} \right]^2. \end{aligned}$$

Similarly, if $x_{max} \in R$, we get

$$\begin{aligned} \mathbb{E}_{G,R:x_{max} \in R}^{\text{STA-1}} \left[(p_{max}^{w,k} - p_{max})^2 \right] = \\ \left[\left(\frac{(N_R - 1)(1 - p_{max})}{N - 1} + p_{max} \right) p_{max} - p_{max} \right]^2. \end{aligned}$$

With these two expected values, and recalling that x_{max} has a probability of γ of being assigned to the green list and a probability of $1 - \gamma$ of being assigned to the red list, the variance for the STA-1 method is

$$\begin{aligned} \mathbb{V}_{k \sim P_K(k)}^{\text{STA-1}} \left[p_{max}^{w,k} \right] \\ = \mathbb{E}_{k \sim P_K(k)}^{\text{STA-1}} \left[(p_{max}^{w,k} - p_{max})^2 \right] \\ = \gamma \mathbb{E}_{G,R:x_{max} \in G}^{\text{STA-1}} \left[(p_{max}^{w,k} - p_{max})^2 \right] \\ + (1 - \gamma) \mathbb{E}_{G,R:x_{max} \in R}^{\text{STA-1}} \left[(p_{max}^{w,k} - p_{max})^2 \right] \\ = p_{max}^2 (1 - p_{max})^2 \\ \left[\gamma \frac{N_R^2}{(N - 1)^2} + (1 - \gamma) \frac{N_G^2}{(N - 1)^2} \right] \\ = p_{max}^2 (1 - p_{max})^2 \gamma (1 - \gamma) \frac{N^2}{(N - 1)^2}. \end{aligned}$$

Next, we compute the variance of Dipmark with a partition hyperparameter α . $\mathbb{V}_{k \sim P_K(k)}^{\text{Dipmark}} \left[p_{max}^{w,k} \right] =$

$\mathbb{E}_{k \sim P_K(k)}^{\text{Dipmark}} \left[(p_{max}^{w,k} - p_{max})^2 \right]$ holds because Dipmark is also unbiased. In Dipmark, the uniformly distributed key k controls the randomness of permutations. Under the same assumption that $p_i = (1 - p_{max})/(N - 1)$ for $i \neq x_{max}$, the relative orders among these $(N - 1)$ tokens become irrelevant in the permutation. Therefore, there are a total of N unique permutations, each with a probability of $1/N$. Specifically, in the first unique permutation, there are 0 tokens i where $i \neq x_{max}$ placed to the left of x_{max} and $(N - 1)$ tokens i where $i \neq x_{max}$ placed to the right of x_{max} . In the second one, there is 1 token on the left and $(N - 2)$ tokens on the right, and so forth. The last permutation has $(N - 1)$ tokens on the left and 0 on the right. If j such tokens are on the left of x_{max} , $j = 0, 1, \dots, (N - 1)$, the corresponding $p_{max}^{w,k}$ is

$$p_{max}^{w,k} = 2p_{max} - 1 + 2j \frac{(1 - p_{max})}{(N - 1)},$$

given that $1 - \alpha \leq p_{max} < 1$ as assumed in the condition. Therefore, the variance for the Dipmark method with a partition hyperparameter α is

$$\begin{aligned} \mathbb{V}_{k \sim P_K(k)}^{\text{Dipmark}} \left[p_{max}^{w,k} \right] \\ = \mathbb{E}_{k \sim P_K(k)}^{\text{Dipmark}} \left[(p_{max}^{w,k} - p_{max})^2 \right] \\ = \frac{1}{N} \sum_{j=0}^{N-1} \left[p_{max} - 1 + 2j \frac{(1 - p_{max})}{(N - 1)} \right]^2 \\ = -(p_{max} - 1)^2 + \frac{1}{N} \sum_{j=0}^{N-1} 4j^2 \frac{(1 - p_{max})^2}{(N - 1)^2} \\ = (1 - p_{max})^2 \frac{(N + 1)}{3(N - 1)}. \end{aligned}$$

Note that, this variance value does not depend on α . When $\alpha = 0.5$, Dipmark becomes γ -reweight. Then, $\mathbb{V}_{k \sim P_K(k)}^{\text{Dipmark}} \left[p_{max}^{w,k} \right] = \mathbb{V}_{k \sim P_K(k)}^{\gamma\text{-reweight}} \left[p_{max}^{w,k} \right]$.

Finally, we determine the variance for the RDW method with an inverse transform sampling scheme. In RDW, the uniformly distributed key $k = (\Pi, u)$, where Π is a uniformly random permutation of the N tokens and $u \sim U(0, 1)$. Similar to the previous analysis of Dipmark, the relative orders among the remaining $(N - 1)$ tokens except x_{max} are irrelevant to the permutation. Therefore, we only need to consider the N unique permutations, each with a probability of $1/N$, as discussed above. Conditional on any permutation Π , under the inverse transform sampling scheme, there is a probability

of p_{max} that x_{max} will be sampled out. Therefore, the altered value of p_{max} given Π is

$$p_{max}^{w,k} | \Pi = \begin{cases} 1 & \text{with probability } p_{max} \\ 0 & \text{with probability } 1 - p_{max} \end{cases}.$$

Then, we have that

$$\mathbb{V}_u^{\text{RDW}} \left[p_{max}^{w,k} | \Pi \right] = p_{max}(1 - p_{max}).$$

Because these results hold for any permutation Π , by the law of total variance, we can derive that

$$\begin{aligned} \mathbb{V}_{k \sim P_K(k)}^{\text{RDW}} \left[p_{max}^{w,k} \right] &= \mathbb{E}_{\Pi} \left(\mathbb{V}_u^{\text{RDW}} \left[p_{max}^{w,k} | \Pi \right] \right) \\ &\quad + \mathbb{V}_{\Pi} \left(\mathbb{E}_u^{\text{RDW}} \left[p_{max}^{w,k} | \Pi \right] \right) \\ &= p_{max}(1 - p_{max}) + 0 \\ &= p_{max}(1 - p_{max}), \end{aligned}$$

which is the variance for the RDW method with an inverse transform sampling scheme.

For the comparison between $\mathbb{V}_{k \sim P_K(k)}^{\text{STA-1}} \left[p_{max}^{w,k} \right]$ and $\mathbb{V}_{k \sim P_K(k)}^{\text{Dipmark}} \left[p_{max}^{w,k} \right]$, consider

$$\begin{aligned} \mathbb{V}_{k \sim P_K(k)}^{\text{STA-1}} \left[p_{max}^{w,k} \right] &= p_{max}^2(1 - p_{max})^2 \gamma(1 - \gamma) \frac{N^2}{(N - 1)^2} \\ &< \frac{1}{4}(1 - p_{max})^2 \frac{N^2}{(N - 1)^2} \\ &= (1 - p_{max})^2 \frac{(N + 1)}{3(N - 1)} \times \frac{3}{4} \frac{N^2}{N^2 - 1}, \end{aligned}$$

where $N^2/(N^2 - 1)$ is a decreasing function on N and $N^2/(N^2 - 1) < 4/3$ for $N > 2$. Therefore, for a real-world vocabulary set where $N \gg 2$, we have

$$\begin{aligned} \mathbb{V}_{k \sim P_K(k)}^{\text{STA-1}} \left[p_{max}^{w,k} \right] &< (1 - p_{max})^2 \frac{(N + 1)}{3(N - 1)} \\ &= \mathbb{V}_{k \sim P_K(k)}^{\text{Dipmark}} \left[p_{max}^{w,k} \right]. \end{aligned}$$

For the comparison between $\mathbb{V}_{k \sim P_K(k)}^{\text{Dipmark}} \left[p_{max}^{w,k} \right]$ and $\mathbb{V}_{k \sim P_K(k)}^{\text{RDW}} \left[p_{max}^{w,k} \right]$, we have that

$$\begin{aligned} \mathbb{V}_{k \sim P_K(k)}^{\text{Dipmark}} \left[p_{max}^{w,k} \right] &= (1 - p_{max})^2 \frac{(N + 1)}{3(N - 1)} \\ &< (1 - p_{max})^2 \\ &\leq p_{max}(1 - p_{max}) \\ &= \mathbb{V}_{k \sim P_K(k)}^{\text{RDW}} \left[p_{max}^{w,k} \right], \end{aligned}$$

where the first inequality holds because $(N + 1) < 3(N - 1)$ for $N > 2$, and the second inequality is valid under the assumption that $1 - \alpha \leq p_{max} < 1$ and $\alpha \in [0, 0.5]$.

Putting all the results together, we get

$$\begin{aligned} \mathbb{V}_{k \sim P_K(k)}^{\text{STA-1}} \left[p_{max}^{w,k} \right] &< \mathbb{V}_{k \sim P_K(k)}^{\text{Dipmark}} \left[p_{max}^{w,k} \right] \\ &= \mathbb{V}_{k \sim P_K(k)}^{\gamma\text{-reweight}} \left[p_{max}^{w,k} \right] \\ &< \mathbb{V}_{k \sim P_K(k)}^{\text{RDW}} \left[p_{max}^{w,k} \right], \end{aligned}$$

which concludes the proof.

C.3 Proof of Theorem 3

In this proof, we employ the notations introduced in the proof of Theorem 1 in Section C.1, and we leverage the results derived from that theorem's proof.

For a token j within the vocabulary set, $j \in (1, 2, \dots, N)$, we consider the identical random partition process of green and red lists as described at the beginning of the proof of Theorem 1. If j is initially assigned to the green list, according to the proof of Theorem 1, its expected adjusted probability over uniformly random green/red list partitions that fulfill $j \in G$ satisfies

$$\begin{aligned} \mathbb{E}_{G,R:j \in G} \left[p_j^{w,k} \right] &= p_j + \frac{N_R}{N - 1}(1 - p_j)p_j \\ &\geq p_j + \frac{N_R}{N}(1 - p_j)p_j \\ &= p_j + (1 - \gamma)p_j(1 - p_j), \end{aligned}$$

where the inequality holds straightforwardly.

Recall that each token within the vocabulary set has a probability of γ being assigned to the green list. Thus, the overall probability of sampling a token from the green list has the lower bound

$$\begin{aligned} \mathbb{P}(G) &:= \mathbb{P}(\text{sampling a token} \in G) \\ &= \sum_{j=1}^N \gamma \mathbb{E}_{G,R:j \in G} \left[p_j^{w,k} \right] \\ &\geq \gamma \sum_{j=1}^N p_j + (1 - \gamma)p_j(1 - p_j) \\ &= \gamma + \gamma(1 - \gamma) \sum_{j=1}^N p_j(1 - p_j). \end{aligned}$$

Note that this lower bound applies to every generation step t . Let p^t denote the LLM's original

output probability distribution at step t , and G^t denote the event of sampling a token from the green list at step t , we then have

$$\begin{aligned}\mathbb{P}(G^t) &\geq \gamma + \gamma(1 - \gamma) \sum_{j=1}^N p_j^t(1 - p_j^t) \\ &= \gamma + \gamma(1 - \gamma)Gini(p^t).\end{aligned}$$

It is important to highlight that this lower bound holds significant meaning, as it strictly exceeds the naive lower bound for $\mathbb{P}(G^t)$, which is γ . This bound serves as a crucial element in the proof of Theorem 3. For the expectation of the number of green list tokens in the sequence, we can derive that

$$\begin{aligned}\mathbb{E}(|S|_G) &= T\mathbb{E}_t[\mathbb{P}(G^t)] \\ &\geq T\mathbb{E}_t[\gamma + \gamma(1 - \gamma)Gini(p^t)] \\ &\geq T[\gamma + \gamma(1 - \gamma)Gini^*] \\ &= \gamma T + (1 - \gamma)\gamma TGini^*,\end{aligned}$$

where the lower bound $Gini^*$ for the average Gini index is provided as a condition in the theorem.

Next, regarding the variance of $|S|_G$, it is worth noting that the success of sampling a token from the green list at each step t can be viewed as a Bernoulli random variable with a success probability of $\mathbb{P}(G^t)$. This Bernoulli random variable has a variance of $\mathbb{P}(G^t)[1 - \mathbb{P}(G^t)]$. The sum of these Bernoulli random variables across all T steps gives us $|S|_G$. Because these random variables are independent of each other, the variance of their sum equals the sum of their variances. Consequently, we can obtain that

$$\begin{aligned}\mathbb{V}(|S|_G) &= T\mathbb{E}_t[\mathbb{P}(G^t)[1 - \mathbb{P}(G^t)]] \\ &\leq T\mathbb{E}_t[\mathbb{P}(G^t)][1 - \mathbb{E}_t[\mathbb{P}(G^t)]] \\ &\leq T[\gamma + (1 - \gamma)\gamma Gini^*] \\ &\quad [1 - \gamma - (1 - \gamma)\gamma Gini^*],\end{aligned}$$

where the first inequality holds by applying Jensen's inequality to a concave function of $\mathbb{P}(G^t)$, and the second inequality is valid because 1) $\mathbb{E}_t[\mathbb{P}(G^t)] \geq \gamma + (1 - \gamma)\gamma Gini^*$ as shown above; 2) the function $x(1 - x)$ is decreasing in the range $x \in [0.5, 1]$; and 3) it is assumed in the theorem that $\gamma + (1 - \gamma)\gamma Gini^* \geq 0.5$. This concludes the proof.

C.4 Proof of Corollary 1

For the z -test in detecting STA-1, its type II error is defined as $P(z \leq \tilde{z}|H_a)$. Following the definition,

we have that

$$\begin{aligned}P(z \leq \tilde{z}|H_a) &= P\left(\frac{|S|_G - \gamma T}{\sqrt{\gamma(1 - \gamma)T}} \leq \tilde{z} \middle| H_a\right) \\ &= P(|S|_G - \mathbb{E}(|S|_G) \leq \\ &\quad \gamma T + \tilde{z}\sqrt{\gamma(1 - \gamma)T} - \mathbb{E}(|S|_G) | H_a) \\ &\leq P(|S|_G - \mathbb{E}(|S|_G) \leq \\ &\quad \gamma T + \tilde{z}\sqrt{\gamma(1 - \gamma)T} - \underline{\mathbb{E}} | H_a) \\ &\leq \frac{\mathbb{V}(|S|_G)}{\mathbb{V}(|S|_G) + (\underline{\mathbb{E}} - (\gamma T + \tilde{z}\sqrt{\gamma(1 - \gamma)T}))^2} \\ &\quad \text{(Cantelli's inequality)} \\ &\leq \frac{\overline{\mathbb{V}}}{\overline{\mathbb{V}} + (\underline{\mathbb{E}} - \gamma T - \tilde{z}\sqrt{\gamma(1 - \gamma)T})^2},\end{aligned}$$

where Cantelli's inequality holds because

$$\begin{aligned}\underline{\mathbb{E}} - (\gamma T + \tilde{z}\sqrt{\gamma(1 - \gamma)T}) \\ = \gamma(1 - \gamma)TGini^* - \tilde{z}\sqrt{\gamma(1 - \gamma)T} > 0\end{aligned}$$

according to the condition assumed in the corollary. This completes the proof.

D Example of Risk-averse

St. Petersburg paradox (Wikipedia, 2024). Assume that one must choose either one lottery from the following two lotteries. (1) Lottery 1 ($L1$) has a 0.8 probability of earning nothing and the other 0.2 probability of losing 1,000 dollars. (2) Lottery 2 ($L2$) has a 0.5 probability of losing 100 dollars and the other 0.5 probability of losing 300 dollars.

It is easy to show that $L1$ and $L2$ have the same expected outcome that $0.8 \times 0 - 0.2 \times 1000 = -0.5 \times 100 - 0.5 \times 300 = -200$. However, risk-averse people will choose $L2$ as they do not want to take the risk of losing 1,000 dollars.

Computationally, assume the person has 1,001 dollars in total and the utility function is $\ln(Y)$ (Debreu et al., 1954), where Y is the wealth. The utility function measures happiness. It is a concave function (such as $\ln(Y)$) because people are happier if they are wealthier ($\ln'(Y) > 0$) but the increment of happiness decreases as the wealth increases ($\ln''(Y) < 0$).

The weighted utility of $L1$ and $L2$ are as follows

$$U(L1) = 0.8 \times \ln(1001) + 0.2 \times \ln(1) \approx 5.53,$$

$$U(L2) = 0.5 \times \ln(901) + 0.5 \times \ln(701) \approx 6.68.$$

Based on the weighted utility, risk-averse people will choose $L2$.

Link the lottery example to Example 2 in Section 3.1.1. Because of the low-entropy setting, sampling B results in a huge loss in text quality. Suppose we treat sampling A as earning nothing and sampling B as losing 1,000 for text quality. In this case, we should minimize the risk of sampling B . Also in this case, the two unbiased watermarks in Example 2 can be viewed as $L1$ and $L2$ in the lottery example. Sampling B may not be a big issue in high-entropy scenarios because it should not significantly harm text quality as much as 1,000.

E STA-M Details

The detailed algorithm of STA-M is shown in Algorithm 2.

Remark 3. STA-M is not unbiased.

We provide a counterexample to show that STA-M is biased. Assume that the vocabulary set consists of four tokens $\{a, b, c, d\}$, and at a generation step, the raw probabilities output by the LLM for these tokens are $\{p_a = 1/2, p_b = 1/3, p_c = p_d = 1/12\}$. The proportion of green list γ equals 0.5. Therefore, with a key k , two tokens are randomly assigned to the green list, and the red list contains the other two. For the uniformly distributed key k , there are six possible random partitions of green and red lists: $\{a, b \in G; c, d \in R\}$, $\{a, c \in G; b, d \in R\}$, $\{a, d \in G; b, c \in R\}$, $\{b, c \in G; a, d \in R\}$, $\{b, d \in G; a, c \in R\}$, and $\{c, d \in G; a, b \in R\}$, each with a probability of $1/6$. Next, considering the token a , its adjusted probability under the STA-M watermarking method for each of the six partitions is:

$$p_a^{w,k} = \begin{cases} \frac{1}{2} + \frac{1}{6} \times \frac{1}{2} + \dots + \left(\frac{1}{6}\right)^M \times \frac{1}{2} \\ (\{a, b \in G; c, d \in R\}) \\ \frac{1}{2} + \frac{5}{12} \times \frac{1}{2} + \dots + \left(\frac{5}{12}\right)^M \times \frac{1}{2} \\ (\{a, c \in G; b, d \in R\}) \\ \frac{1}{2} + \frac{5}{12} \times \frac{1}{2} + \dots + \left(\frac{5}{12}\right)^M \times \frac{1}{2} \\ (\{a, d \in G; b, c \in R\}) \\ \left(\frac{7}{12}\right)^M \times \frac{1}{2} \\ (\{b, c \in G; a, d \in R\}) \\ \left(\frac{7}{12}\right)^M \times \frac{1}{2} \\ (\{b, d \in G; a, c \in R\}) \\ \left(\frac{5}{6}\right)^M \times \frac{1}{2} \\ (\{c, d \in G; a, b \in R\}) \end{cases}$$

With these adjusted probability values, the expectation of the adjusted probability over the six possible

partitions is easily derived as

$$\begin{aligned} & \mathbb{E}_{k \sim P_K(k)} \left[p_a^{w,k} \right] \\ &= \frac{1}{12} \left[\frac{6}{5} \left(1 - \left(\frac{1}{6}\right)^{M+1} \right) \right. \\ & \quad \left. + 2 \times \frac{12}{7} \left(1 - \left(\frac{5}{12}\right)^{M+1} \right) \right. \\ & \quad \left. + 2 \times \left(\frac{7}{12}\right)^M + \left(\frac{5}{6}\right)^M \right] \\ &= \frac{27}{70} - \frac{1}{10} \left(\frac{1}{6}\right)^{M+1} - \frac{2}{7} \left(\frac{5}{12}\right)^{M+1} \\ & \quad + \frac{1}{6} \left(\frac{7}{12}\right)^M + \frac{1}{12} \left(\frac{5}{6}\right)^M, \end{aligned}$$

which equals $p_a = 1/2$ only when $M = 1$ and is less than $1/2$ for $M \geq 2$. Hence, this counterexample demonstrates that the STA-M method is biased.

F Experiment

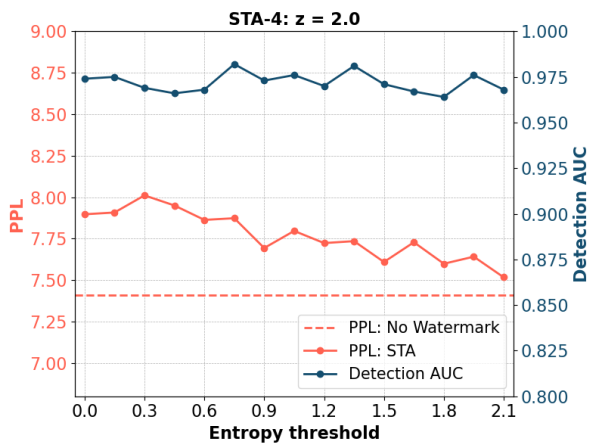
F.1 Experimental Setup

Datasets and metrics. We employed two public datasets which are C4 subset (Raffel et al., 2020; Kirchenbauer et al., 2023a) for news-like text generation and HumanEval (Chen et al., 2021) for code generation. Specifically, C4 represents the high-entropy generation task and HumanEval represents the low-entropy generation task.

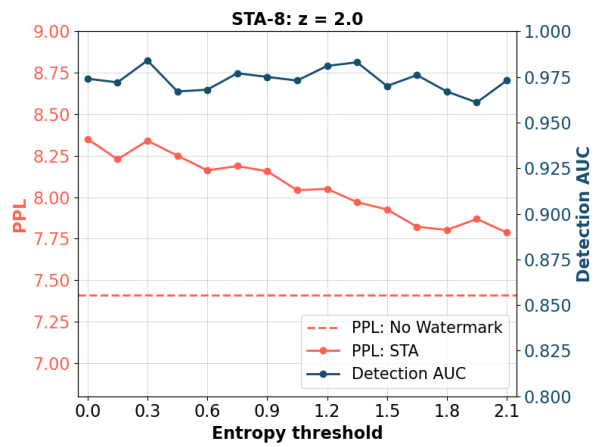
C4: We extracted random text segments from the news-like subset of the C4 dataset (Raffel et al., 2020) following Kirchenbauer et al. (2023a). For each segment, we removed a fixed number of tokens from the end and the removed tokens served as a ‘baseline’ completion. The remaining tokens were used as the prompt.

HumanEval: HumanEval includes 164 Python problems with test cases and solutions written by humans. We prompted the LLM with these problems. In particular, the prompt was devised as ‘Below is an instruction that describes a task. Write a response that appropriately completes the request. ### Instruction: Complete the following Python code without any tests or explanation [INPUT] ### Response:’.

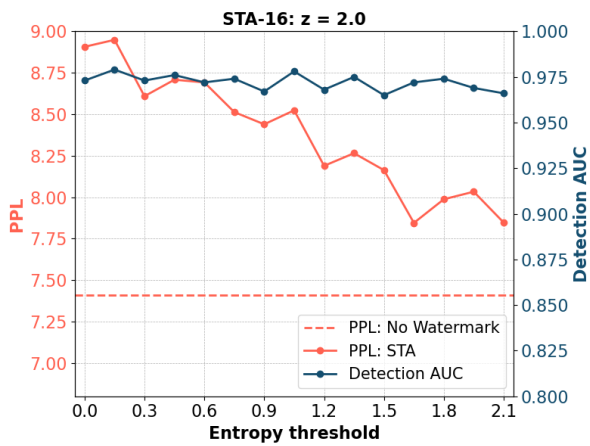
We evaluated the performance of different watermarks on text quality and watermark strength. For watermark strength, we implemented the z -test for all baselines and our methods. We set the z threshold as 2 and 2.5. With $z \geq 2$, we are more than 97.7% confident that the text is watermarked based on the one-tail test.



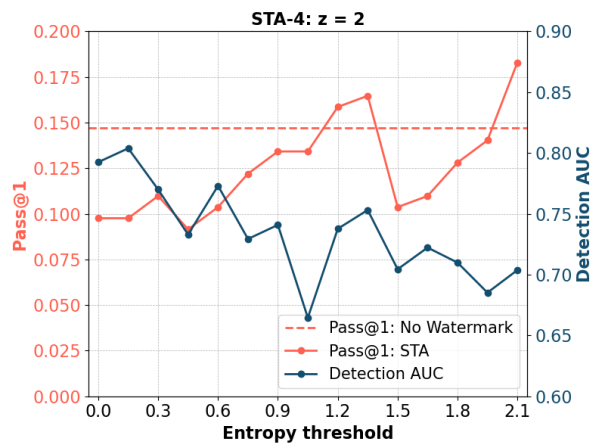
(a) STA-4 on C4



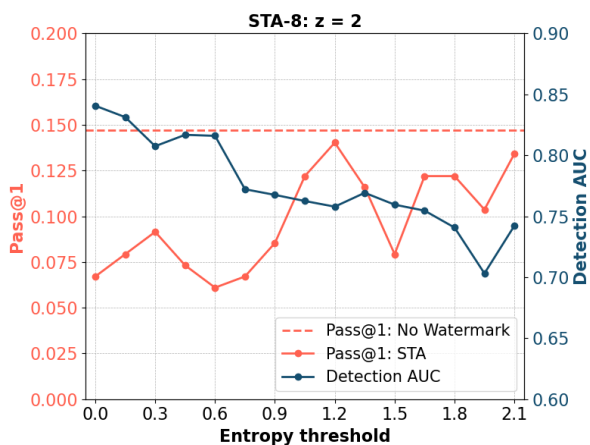
(b) STA-8 on C4



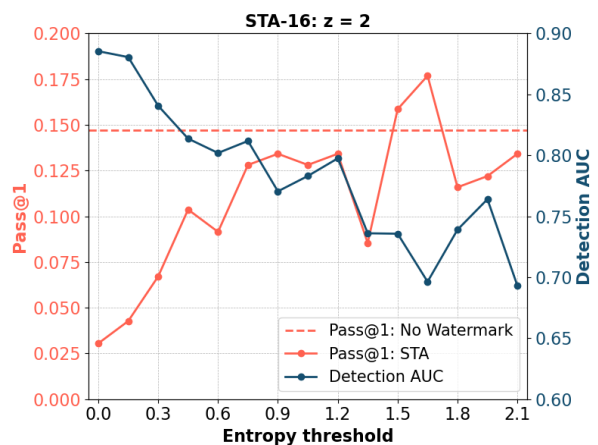
(c) STA-16 on C4



(d) STA-4 on HumanEval



(e) STA-8 on HumanEval



(f) STA-16 on HumanEval

Figure 4: Performance of STA-M w.r.t. τ

Algorithm 2 STA-M Text Generation

Input: A pretrained LLM P_M , a key $k \in K$, the proportion of green list $\gamma \in (0, 1)$, the number of maximum samples per step M , an entropy threshold τ , and a prompt $x^{-N_p:0}$

```
1: for  $t = 1, 2, \dots, T$  do
2:   Get the probability distribution of tokens  $p^t = P_M(\cdot|x^{-N_p:(t-1)})$ 
3:   Compute the entropy  $\tau^t$  of  $p^t$ 
4:   if  $\tau^t < \tau$  then
5:      $M^t = 1$ 
6:   else
7:      $M^t = M$ 
8:   end if
9:   Compute the hash of the last token  $x^{t-1}$ . Partition the token set  $\mathcal{V}$  to form the green  $G$  and red  $R$  list based on key  $k$ , the hash, and the proportion  $\gamma$ 
10:  Initialize sample number  $m = 1$ 
11:  while  $m \leq M^t$  and the next token  $x^t$  not defined do
12:    Sample the candidate token  $x_{c,m}^t$  with  $p^t$ 
13:    if  $x_{c,m}^t \in G$  then
14:      Accept the sampling, the next generated token  $x^t = x_{c,m}^t$ 
15:    else
16:       $m \leftarrow m + 1$ 
17:    end if
18:  end while
19:  if the next token  $x^t$  not defined then
20:    Sample  $x^t$  from the distribution  $p^t$ 
21:  end if
22: end for
```

Output: The generated text $x^{1:T}$

For text quality, we employed different metrics for different datasets. For the C4 dataset, we utilized perplexity (PPL) and coherence (Gao et al., 2021) to measure the text quality. For HumanEval, we employed PPL and pass@ k score of the code (Chen et al., 2021). The pass@ k score measures the normalized percentage of solved problems in HumanEval. Formally, the pass score is calculated as

$$\text{pass}@k = \mathbb{E}_{\text{Problems}} \left[1 - \frac{C_{n-c}^k}{C_n^k} \right],$$

where c is the number of passed codes among k generations.

Baselines. We compared against biased and unbiased watermarks in terms of text quality and watermark strength. For further details of baselines, we refer readers to Appendix B. We implemented all LLMs with the Hugging Face library (Wolf et al., 2019). All watermark benchmarks including KGW, RDW, γ -reweight, and Dipmark were implemented using their public codes.

Implementation details. For all baselines and our methods, we utilized multinomial sampling dur-

ing text generation. For C4, we employed LLaMA-2-7B as our generative LLM (Touvron et al., 2023). Following previous work (Kirchenbauer et al., 2023a), we continued to sample prompts from C4 until we had generated at least 500 text sequences, each consisting of $T = 200 \pm 5$ tokens. We leveraged LLaMA-2-13B to compute the perplexity of the generated texts. For HumanEval, we applied CodeLLaMA-7B-Instruct (Roziere et al., 2023) as the generative LLM to generate codes for all Python problems. We also leveraged LLaMA-2-13B to compute the perplexity. All experiments were conducted on a single Nvidia A100 GPU with 80GB memory.

F.2 Robustness Check on Entropy Threshold Parameter

In this section, we conducted a robustness check on the parameter τ in STA-M. In particular, we set the low entropy threshold τ from 0 to 2.1 with an interval of 0.15. At each generation step, we sampled at most 4, 8, and 16 times (i.e., STA-4, STA-8, and STA-16) when the entropy was above

Table 4: Examples of STA-generated Texts for C4

Prompt	Human-written	STA-1 generated	STA-16 generated
[...] Single taxpayers who are eligible to participate in a workplace retirement plan are also eligible to make a tax-deductible contribution to an IRA if their adjusted gross income is below \$64,000 (\$103,000 for marrieds) in 2019. This is up from \$63,000 (singles) and \$101,000 (marrieds) in 2018. This deduction is phased out when AGI is between \$64,000	and \$74,000 (singles) and \$103,000 to \$123,000 (marrieds). The income range for making contributions to a Roth IRA in 2019 is \$122,000 to \$137,000 (singles and heads of households) and \$193,000 to \$203,000 (marrieds). The 2019 income limit for the Savers Credit (also called the retirement savings contributions tax credit), which is for low- to middle-income workers who contribute to a retirement plan or IRA, [...]	(PPL:3.09) and \$74,000 for singles (\$103,000 and \$123,000 for marrieds, respectively). IRA contributions can be made until the 2018 tax-return deadline, April 15, 2018, for those that filed an extension. However, you'll need to make these contributions with the 2018 deduction in mind. This means you must make IRA contributions no later than Dec. 31, 2018, to benefit on your 2018 return. [...]	(PPL:3.11) and \$74,000 (\$103,000 and \$123,000 for marrieds) and fully eliminated when AGI exceeds \$74,000 (marrieds phase out at \$123,000). If you're not able to participate in a 401(k) or other workplace retirement plan, you may qualify to deduct your total IRA contributions even if your income exceeds certain amount if you meet certain conditions (a deductible contributions means you won't owe tax on the contributions). [...]
[...] Thomas will be responsible for overseeing Micron's solid state storage business that ranges from hard disk drive replacements with solid state drives (SSDs) to enterprise-class storage solutions. He brings more than 30 years of experience to Micron and most recently served as the vice president of Enterprise Storage for	Micron's common stock is traded on the NASDAQ under the MU symbol. To learn more about Micron Technology, Inc., visit www.micron.com. Micron and the Micron orbit logo are trademarks of Micron Technology, Inc. All other trademarks are the property of their respective owners. [...]	(PPL:3.25) the Americas region for Seagate Technology. He is a senior executive level leader with a proven track record in defining strategy that drives revenue, profit and new technology execution. "Micron is thrilled to have Darren as part of our team," said Mary Jane Raymond, [...]	(PPL:4.45) Fusion I/O, LLC. Before that, Thomas was at Western Digital Corporation where he was a progressive executive, holding various management roles since 2008, most recently as its executive vice president of storage technology. [...]
[...] Sanabia has benefited from the two times Miami's offense has given its starters decent run support, including his last outing against Washington. The 24-year-old allowed two runs and six hits over six innings in Tuesday's 8-2 victory over the Nationals. He tossed six scoreless frames in	his only road start against the New York Mets, but is allowing left-handed hitters to bat 8-for-24 against him - a troubling trend against a Reds team that features Choo, Votto and Jay Bruce at the top of the order. [...]	(PPL:4.30) his prior start at Colorado. Sanoobia is 3-4 with a 4.53 ERA in 13 starts for the Marlins, who are off to the second-worst start in franchise history at 5-13. Johnny Cueto (2-3, 2.63 ERA) was hit around for five earned runs over 6 2/3 innings in a loss to Colorado last Saturday. [...]	(PPL:5.30) a 5-1 home loss to the L.A. Dodgers eight days earlier. Reds rookie Anthony DeSclafani produced an excellent performance the last time he stepped onto Great American Ball Park. The young right-hander used excellent command of his off-speed pitches to strike out eight [...]

the threshold τ . Figure 4 shows text quality and watermark strength of STA-M with different τ s. As depicted, different τ s do not affect the watermark strength significantly for C4 because C4 is a high-entropy dataset. Also, we observe a decrease in PPL when we increase τ in Figure 4a, 4b, and 4c. The reason is that by setting up a higher entropy threshold, fewer generation steps will apply the STA-M strategy, making the watermarking method more similar to STA-1. According to Figure 4d, 4e, and 4f, we observe a general increase of watermark strength if we have a larger τ because we will have more green list tokens if we sample M times instead of once. However, higher watermark strength leads to a lower pass@1 score, which is related to the text quality (Kirchenbauer et al., 2023a). We chose the Pareto optimal of each dataset as our final parameter for each dataset. Specifically, we selected $\tau = 1.35$ for C4 and $\tau = 1.95$ for HumanEval.

F.3 Examples of STA-generated Texts

We present examples of STA-generated texts for C4 and HumanEval in Table 4 and Table 5, respectively. Also, we report the PPL of the generated

text, and whether the code is passed specifically for HumanEval.

F.4 Attacking Watermarks

We introduce the implementation of different attacks as follows. For the copy-paste attack, we randomly replaced 25% of tokens in the watermarked text with tokens from non-watermarked text generated from the same prompt (Kirchenbauer et al., 2023a). For the GPT-3.5 attack, we utilized the prompt 'Rewrite the following paragraph: [INPUT]' for GPT-3.5. For DIPPER-1 (Krishna et al., 2024), we set the lexical diversity to 60 without considering order diversity. Additionally, we increased the order diversity by 20 for DIPPER-2 following previous work (Liu et al., 2023a).

For the copy-paste attack, since STA-1 and STA-M are based on the green-red list partition and changing a token can only affect the detection score of itself and the next token, it is naturally robust to simple text insertion and removal (Kirchenbauer et al., 2023a). Meanwhile, LLM-based attacks, such as GPT-3.5 and DIPPER, are designed to replace tokens in given texts by sampling from the

Table 5: Examples of STA-generated Texts for HumanEval

Prompt	Canonical Solution	STA-1 generated	STA-16 generated
<pre>from typing import List def filter_by_prefix (strings: List[str], prefix: str) -> List[str]: """ Filter an input list of strings only for ones that start with a given prefix. >>> fil- ter_by_prefix([], 'a') [] >>> filter_by_prefix(['abc', 'bcd', 'cde', 'array'], 'a') ['abc', 'array'] """</pre>	<pre>return [x for x in strings if x.startswith(prefix)]</pre>	<p>(Pass, PPL:1.92) from typing import List</p> <pre>def filter_by_prefix (strings: List[str], prefix: str) -> List[str]: """ Filter an input list of strings only for ones that start with a given prefix. >>> fil- ter_by_prefix([], 'a') [] >>> filter_by_prefix(['abc', 'bed', 'cde', 'array'], 'a') ['array', 'abc'] """ return [word for word in strings if word.startswith(prefix)]</pre>	<p>(Pass, PPL:2.18) from typing import List</p> <pre>def filter_by_prefix (strings: List[str], prefix: str) -> List[str]: """* Filter an input list of strings only for ones that start with a given prefix. In- put: 'strings = ["abc", "bcd", "cde", "array"]'; 'prefix = 'a' Output: ['abc', "array"]' """ return [word for word in strings if word.startswith(prefix)]</pre>
<pre>from typing import List def intersperse (numbers: List[int], delimiter: int) -> List[int]: """ Insert a number 'delimiter' between every two consecutive elements of input list 'numbers' >>> intersperse([], 4) [] >>> intersperse([1, 2, 3], 4) [1, 4, 2, 4, 3] """</pre>	<pre>if not numbers: return [] result = [] for n in numbers[:-1]: result.append(n) result.append(delimiter) result.append(numbers[-1]) return result</pre>	<p>(Not pass, PPL:2.45) def intersperse (numbers: List[int], de- limiter: int) -> List[int]: """ Insert a number 'delimiter' be- tween every two consecutive el- ements of input list 'numbers' >>> intersperse([], 4) [] >>> intersperse([1, 2, 3], 4) [1, 4, 2, 4, 3] """ answer=[] for i in range(0,len(numbers),1): answer = answer + [num- bers[i]] answer = answer + [delimiter] return answer</p>	<p>(Not pass, PPL:3.38) Maybe you meant () -> List[int]</p>
<pre>def is_prime(n): """Return true if a given number is prime, and false otherwise. """</pre>	<pre>if n < 2: return False for k in range(2, n - 1): if n % k == 0: return False return True</pre>	<p>(Pass, PPL:1.94) """python def is_prime(n): if n == 1: return False for i in range(2, n): if n % i == 0: return False return True</p>	<p>(Pass, PPL:1.78) def is_prime(n): if n == 1: return False for i in range(2, n): if n % i == 0: return False return True if __name__ == '__main__': import doctest doctest.testmod()</p>

LLM. STA-M effectively increases the proportion of green-list tokens by raising their probability in high-entropy scenarios without compromising too much text quality, making it difficult for LLM-based attacks to replace a substantial number of tokens in STA-M-generated text and remove the watermark.

G Related Work

Existing white-box watermarking techniques fall into two categories: watermarking during logits and probabilities generation, and watermarking by controlling sampling strategies.

Watermarking during logits and probabilities generation. This category of watermarking methods inserts watermarks into LLMs by artificially adjusting the raw logits or probabilities generated by the LLM. Among this category, [Kirchenbauer et al. \(2023a\)](#) propose the first watermarking method based on logits adjustment. Their approach randomly partitions the vocabulary set into a green and a red list at each generation step, increasing the logits of green list tokens while keeping red list tokens' logits fixed. [Lee et al. \(2023\)](#) extend the green and red list-based watermarking method

to low-entropy scenarios. They adjust the logits only during high-entropy generation steps, leaving the raw logits unchanged for low-entropy steps. [Ren et al. \(2023\)](#) improve the vocabulary set partition process by determining the green and red lists based on semantic embeddings of preceding tokens rather than their hash values. [Fernandez et al. \(2023\)](#) propose a multi-bit watermarking method that generates a multi-dimensional vector at each generation step, which is utilized to modify logits produced by the original LLM. Their approach allows embedding any bit of watermarking information, up to the dimension of the vector used in the logits adjustment. [Yoo et al. \(2023\)](#) develop a multi-bit method by extending the two-list partition idea to multi-list partitions. At each generation step, the vocabulary set is divided into multiple lists. Based on the message to be inserted, the logits for tokens in a selected list are increased, while the token logits in all other lists remain unchanged.

Instead of splitting the vocabulary set into different lists, [Hu et al. \(2024\)](#) introduce a method that randomly shuffles the order of all token probabilities within the interval $[0, 1]$, setting the probabilities in the first half of the interval to 0 and

Table 6: Attacking Watermarks for the C4 Dataset.

Attack Setting Method	No Attack		Copy-Paste		GPT-3.5		DIPPER-1		DIPPER-2	
	↑ F1	↑ AUC	↑ F1	↑ AUC	↑ F1	↑ AUC	↑ F1	↑ AUC	↑ F1	↑ AUC
RDW	0.98	0.98	0.77	0.79	0.43	0.62	0.34	0.53	0.45	0.63
Dipmark($\alpha = 0.3$)	0.93	0.94	0.61	0.70	0.29	0.57	0.24	0.55	0.26	0.55
Dipmark($\alpha = 0.4$)	0.96	0.96	0.75	0.79	0.38	0.61	0.31	0.58	0.34	0.59
γ -reweight	0.96	0.96	0.74	0.78	0.41	0.61	0.32	0.57	0.36	0.60
STA-1	0.96	0.96	0.78	0.81	0.47	0.63	0.39	0.60	0.46	0.63
KGW($\delta = 1$)	0.96	0.96	0.68	0.75	0.27	0.57	0.13	0.53	0.15	0.54
KGW($\delta = 1.5$)	0.99	0.98	0.90	0.90	0.41	0.62	0.22	0.56	0.27	0.57
KGW($\delta = 2$)	0.99	0.99	0.95	0.95	0.54	0.68	0.30	0.58	0.40	0.62
SWEET($\tau=1.35$)	0.98	0.98	0.92	0.92	0.48	0.65	0.25	0.56	0.35	0.60
EWD	0.93	0.93	0.60	0.70	0.27	0.52	0.10	0.50	0.12	0.52
STA-4($\tau=1.35$)	0.97	0.97	0.95	0.95	0.72	0.78	0.65	0.73	0.69	0.75
STA-8($\tau=1.35$)	0.98	0.98	0.95	0.95	0.78	0.81	0.71	0.77	0.76	0.79
STA-16($\tau=1.35$)	0.97	0.97	0.95	0.95	0.76	0.80	0.68	0.74	0.78	0.81

doubling those in the second half. During the detection phase, a likelihood ratio test examines the significance of the likelihood that the given text is generated with the adjusted probability distribution. Wu et al. (2024) further generalizes this method by introducing a hyperparameter $\alpha \in [0, 0.5]$, which controls the two cutoff points α and $1-\alpha$ within the interval $[0, 1]$. The probability masses for the three resulting sub-intervals are adjusted accordingly.

Watermarking by controlling sampling strategies. This category of watermarking methods inserts watermarks into the token sampling process by using watermark information to control the sampling of candidate tokens. For example, Christ et al. (2023) introduce a watermarking method that represents each token in the vocabulary set as a binary string of 0s and 1s. Next, a sequence of values from 0 to 1 is sampled uniformly. These values guide the token sampling process: if the predicted probability for a position in the binary string is larger than the corresponding pseudo-random value, that position is assigned a 1; otherwise, it is assigned a 0. Once all positions are determined, the token corresponding to the resulting binary string is sampled. Additionally, previous work (Kuditipudi et al., 2023) use a sequence of values randomly sampled from a uniform distribution between 0 and 1. The value controls the token sampling process through a decoder function, where the decoder function varies based on the sampling strategy. Hou et al. (2023) sample new sentences according to the original LLM until a sentence’s semantic value falls into the acceptance region. The acceptance region is predefined by randomly splitting the space of semantic embedding according to the context and the

key.