# Low-Entropy Watermark Detection via Bayes' Rule Derived Detector

**Beining Huang**[1,2], **Du Su**[1†], **Fei Sun**[1], **Qi Cao**[1], **Huawei Shen**[1,2], **Xueqi Cheng**[1,2]

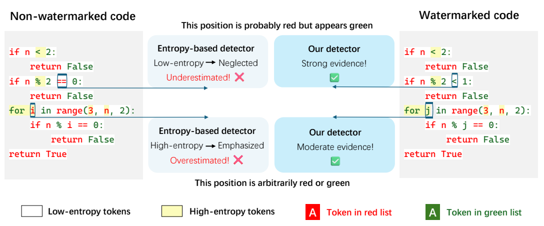
[1]State Key Laboratory of AI Safety,
Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China
[2]University of Chinese Academy of Sciences, Beijing, China
`{huangbeining23s, sudu, sunfei, caoqi, shenhuawei, cxq}@ict.ac.cn`

## Abstract

Text watermarking, which modify tokens to embed watermark, has proven effective in detecting machine-generated texts. Yet its application to low-entropy texts like code and mathematics presents significant challenges. A fair number of tokens in these texts are hardly modifiable without changing the intended meaning, causing statistical measures to falsely indicate the absence of a watermark. Existing research addresses this issue by rely mainly on a limited number of high-entropy tokens, which are considered flexible for modification, and accurately reflecting watermarks. However, their detection accuracy remains suboptimal, as they neglect strong watermark evidences embedded in low entropy tokens modified through watermarking. To overcome this limitation, we introduce **B**ayes' **R**ule derived **W**atermark **D**etector (BRWD), which exploit watermark information from every token, by leveraging the posterior probability of watermark's presence. We theoretically prove the optimality of our method in terms of detection accuracy, and demonstrate its superiority across various datasets, models, and watermark injection strategies. Notably, our method achieves up to 50% and 70% relative improvements in detection accuracy over the best baselines in code generation and math problem-solving tasks, respectively.

## 1 Introduction

Text watermarking is an effective technique for differentiating machine-generated text from human-written content that subtly injects an invisible marker, i.e. watermark, into text. It serves as a safeguard against unauthorized or malicious use of large language models, such as creating fake news (Augenstein et al., 2023) and election manipulating (Alvarez et al., 2023).



```
if n < 2:
    return False
if n % 2 == 0:
    return False
for i in range(3, n, 2):
    if n % i == 0:
        return False
return True
```

```
if n < 2:
    return False
if n % 2 < 1:
    return False
for j in range(3, n, 2):
    if n % j == 0:
        return False
return True
```

Figure 1: Comparison of entropy-based detector and ours. It demonstrates the misalignment between entropy and informativeness for watermark detection, where low-entropy positions with critical information are overlooked, and high-entropy positions are overestimated.

Generative watermark, which integrates watermark during LLM's generation process, generally exhibits superior detectability and robustness. A generative watermarking method injects a watermark through perturbing tokens' output distribution during text generation, and subsequently computes a score indicating its presence for detection. For instance, the KGW method (Kirchenbauer et al., 2023) partitions the model's vocabulary into green and red tokens at each generation step, and increases the output probability of green tokens, resulting in higher proportion of green tokens in a watermarked text. Subsequent detection is performed by computing the z-score of green token occurrences in the text.

However, text watermarking exhibits suboptimal performance in low-entropy scenarios, such as code generation or mathematical problem solving, where a fair number of tokens are unmodifiable without compromising output quality. Statistics, such as the z-score computed for these tokens, suggest the absence of a watermark, providing contrary evidence of its presence and thus diminishing the effectiveness of the watermark detector. This limitation poses a substantial barrier to detecting malicious or unauthorized activities in software de-

†Corresponding author.
[1]Our Code is available at `https://github.com/cczslp/BRWD`.

velopment, academic exams (Susnjak, 2022), and job interviews (Canagasuriam and Lukacik, 2024), raising concerns about social equality and ethical use of technology. Subsequent works (Lee et al., 2024; Lu et al., 2024) attempt to enhance detection accuracy by prioritizing high-entropy tokens, which are considered more adjustable to watermark perturbations, thus more indicative of watermark's presence. In contrast, low-entropy tokens receive little weight during detection. Nevertheless, their detection accuracy remains suboptimal, as they fail to leverage the substantial information in highly indicative low-entropy tokens, which are actually modified through watermarking.

We identified a ***misalignment*** between the entropy-based mechanism and the goal of detection: Entropy inadequately captures the modifiability of a token, nor does it account for the actual modification introduced by watermarking. To illustrate the misalignment, consider a low-entropy position where the probability is concentrated on a red token but presents a green one. It serves as a strong evidence of the watermark's presence. However, entropy-based detectors would assign a small weight to this position and neglect this evidence.

To address this misalignment, we propose **B**ayes' **R**ule-derived-**W**atermark **D**ector (BRWD), which quantifies the impact of watermark injection on every token. We calculate the posterior likelihood of every token altered by watermark injection with Bayes' rule, then aggregate it into a total score. A token that deviates more from its original distribution and consequently shows stronger evidence of watermark injection has a greater impact on the score. In this way, we extract more information from every token, especially highly indicative and low-entropy tokens, instead of neglecting them.

We prove that BRWD *achieves optimal* true positive rate (TPR), given any false positive rate (FPR) limit. Our method is compatible with various watermark injection techniques. Extensive experiments across multiple language models, generative tasks and watermark injection schemes, under scenarios with or without prompts, demonstrate the superior performance of our approach. In particular, under a 1% FPR constraint in mathematical contexts, BRWD boosts the TPR from below 60% to over 90%. Furthermore, BRWD demonstrates adaptability to general high-entropy texts and exhibits robustness against removal attacks.

https://www.recordedfuture.com/research/i-chatbot

In summary, our main contributions are:

- We propose a watermark detection approach called BRWD, which significantly improves watermark detection accuracy in low-entropy scenarios.
- We provide theoretical analysis on the optimality of BRWD under any constraint on false positive rate.
- We conduct experiments with various watermarking methods in low-entropy scenarios and empirically verify BRWD's superiority.

## 2 Related Works

Text watermarking emerges as a promising solution to identifying machine-generated contents. They can be categorized into two types (Liu et al., 2024b): generative watermarking methods and watermarking applied to existing text.

**Generative Watermarking Methods.** Generative watermarking methods inject watermarks during the LLM generation phase through modifying model output logits or meticulously designed sampling process.

KGW (Kirchenbauer et al., 2023) is a seminal approach within logits-modifying type, and its proposed green-red list paradigm has been widely adopted by subsequent studies. Observing its impact on the quality of generated text, several studies (Huo et al., 2024; Chen et al., 2024) have proposed improvements. For instance, TS-watermark (Huo et al., 2024) dynamically adjusts the watermark strength for each token based on the preceding token. Some other works enhance the robustness of KGW through using fixed green list (Zhao et al., 2023) or determining the green list by semantics (Liu et al., 2024a; Liu and Bu, 2024). Additionally, some studies (Fernandez et al., 2023; Wang et al., 2024; Yoo et al., 2024) focused on injecting watermark with more information (multi-bit watermark). Research also explored the adaptability of watermarking methods to low-entropy scenarios (Lee et al., 2024; Lu et al., 2024), such as programming tasks. Logits-modifying watermarking methods typically adopt z-score based detector. To improve detectability in low-entropy scenarios, Lu et al. (2024) introduced EWD detector based on token entropy.

Sampling-based watermarking methods introduce pseudo-randomness accessible during detection to influence the sampling. For example, Kuditipudi et al. (2023) employed a long random number sequence to modify sampling and used edit distance for detection to enhance robustness. Dathathri et al. (2024) introduced tournament sampling, achieving a balance between text quality, detectability, and efficiency.

**Watermarking for Existing Texts.** This type of watermarking methods inject hidden features into existing texts and detect them afterwards. One way to achieve this is using end-to-end models (Abdelnabi and Fritz, 2021; Zhang et al., 2024). For instance, AWT (Abdelnabi and Fritz, 2021) employs a transformer network to inject watermarks and another transformer network to detect them. Some other methods attempt to watermark existing texts through synonym substitutions. The techniques they adopt for synonym selection include consulting an electronic dictionary (Topkara et al., 2006) and utilizing pre-trained models (Yang et al., 2023; Yoo et al., 2023). The quality of watermarked text generated by such methods is constrained by the synonym database or model used.

## 3 Preliminaries

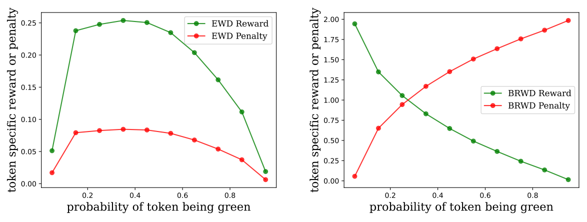
As we focus on generative watermarking due to its superior detectability and robustness, the term "watermarking" will henceforth refer to generative watermarking. In this section, we present the preliminaries of watermark injection and detection. Notations used throughout the paper are provided in Appendix A.

### 3.1 Watermark Injection

Watermark injection is typically achieved by perturbing the output distribution of large language model. Specifically, when generating the $t$-th token, the language model $M$ computes a logits vector $\boldsymbol{l}_t$ from preceding context, which is then normalized via $\mathrm{softmax}$ to produce a distribution over candidate tokens. Watermarking methods introduce perturbations to this distribution by modifying the logits or sampling process.

In this study, we select two representative watermarking approaches as research objects: the classical KGW (Kirchenbauer et al., 2023) method and SWEET (Lee et al., 2024), the state-of-the-art method for watermarking in low-entropy scenarios.

At generation step $t$, KGW adds a bias $\delta$ to the logits of a subset of tokens in vocabulary $\mathcal{V}$, namely green list, denoted by $\mathcal{V}_g$. It is determined by the hash of preceding tokens and the proportion of $\mathcal{V}_g$



(a) EWD's token scoring mechanism

(b) BRWD's token scoring mechanism

Figure 2: Reward and penalty with respect to total probablity of tokens being green. Statistics from 500 texts in MBPP dataset.

in $\mathcal{V}$ is $\gamma$. The remaining tokens in $\mathcal{V}$ are called "red-list", denoted by $\mathcal{V}_r$. This logits perturbation increases the probability of green tokens, resulting in a generated text with more green tokens.

The SWEET method follows the same procedure but restricts perturbations to high-entropy positions only. In SWEET method, the entropy at position $t$ is $H_t = -\sum_{v \in \mathcal{V}} p_t^v \log p_t^v$ where $p_t^v$ is the probability of candidate token $v$ given by $M$ at position $t$. A threshold is manually set to determine high-entropy positions.

### 3.2 Watermark Detection

We present a watermark detection framework encompassing detectors used by KGW and SWEET. Let $\boldsymbol{x} = \{x_0, x_1, \ldots, x_{T-1}\}$ denote the text to be detected. Within this framework, the detector assigns a score to every token $x_t$ in $\boldsymbol{x}$. These scores are then summed and (optionally) normalized by $\mathrm{norm}(\cdot, \cdot)$ to produce a total score. A higher score indicates a greater likelihood that the text has been watermarked. Specifically, the score given to $\boldsymbol{x}$ is:

$$S(\boldsymbol{x}) = \mathrm{norm}(\sum_{t=0}^{T-1} s(\boldsymbol{x}, t), \boldsymbol{x})$$

where $s$ is the score given to token at each position. In practice, a threshold is defined and any text scored higher than it is identified as watermarked.

## 4 Method

In this section, we first illustrate our motivation by examining the limitations of current watermark detection methods and then present our approach.

### 4.1 Motivation

Although EWD and SWEET detectors achieve better accuracy in low-entropy scenarios, their improvement stems from attaching more importance to those high entropy tokens, which are more adjustable to watermark perturbation. However, the goal of watermark detection is to measure the influence of the perturbation. We explain this misalignment from a token scoring perspective.

The EWD detector assigns a positive score as a reward, if $x_t \in \mathcal{V}_g$, and a negative penalty if $x_t \in \mathcal{V}_r$. As for token-level reward or penalty, we have the following ***intuition***: if a position was likely to output a red token according to its original distribution without watermark but a green token appears here. It indicates the influence of watermark and should receive a significant reward. Inversely, a position with high probability of green tokens but appears a red one strongly suggests that watermark is absent, thus should receive a significant penalty.

However, EWD does not fully exploit these critical evidences for determining the presence of the watermark. Figure 2a illustrates the relationship between the reward and penalty values assigned by EWD ($\gamma = 0.25$) and the probability of green tokens. It can be observed that both reward and penalty exhibit a bell-shaped curve trend. The reward declines when the probability of a token being green is relatively low, and the penalty declines when this probability is relatively high, which contradicts our previous intuition. Other entropy-based methods like SWEET have the same problem.

To further analyze the impact of this misalignment on token scores, we define a watermark information score (WIS). For the $t$-th token $x_t$ in a text $\boldsymbol{x}$, its WIS is:

$$\mathrm{WIS}(\boldsymbol{x}, t) = \mathbf{1}_{\mathcal{V}_g}(x_t) \sum_{v \in \mathcal{V}_r} p_t^v + \mathbf{1}_{\mathcal{V}_r}(x_t) \sum_{v \in \mathcal{V}_g} p_t^v$$

This score quantifies the amount of information a token carries about the watermark. Specifically, for a green token, its WIS equals the total probability of red tokens. A higher value indicates a greater likelihood of the watermark's presence, as the position would otherwise likely contain a red token. For a red token, its WIS equals the total probability of green tokens. A higher value suggests a lower likelihood of the watermark's presence.

We plotted a scatter diagram of the entropy and WIS of tokens in watermarked codes. In Figure 3, the green tokens within the black rectangle exhibit both high WIS and low entropy. This indicates that entropy-based detectors assign these tokens relatively low importance, thereby impairing detection performance. This observation further supports our hypothesis regarding the misalignment in entropy-based detectors.

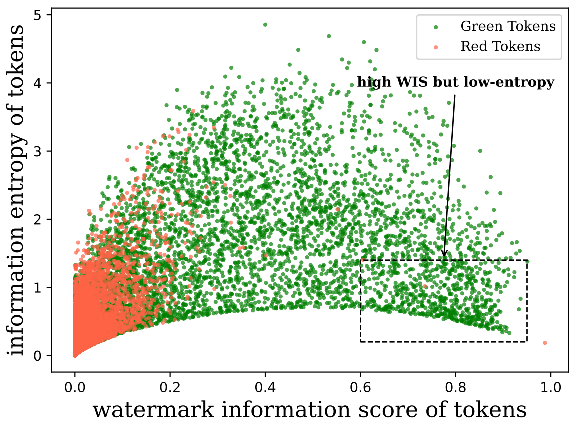


Figure 3: Watermark information score (WIS) and entropy of tokens in watermarked codes. Statistics from 500 watermarked codes generated using MBPP dataset and StarCoder2-7B model.

To address this issue, we devise a detection method based on Bayes' rule measuring watermark influence. We treat the target text $\boldsymbol{x}$, as a random variable, and infer the likelihood of the watermark's presence from it.

## 4.2 Bayes' Rule Derived Detection

Let $w$ denote the presence of a watermark (0/1 binary) and $P_w$ represent the distribution of $w$. Suppose the distribution of all natural language (watermarked and non-watermarked) is $P$ and the distribution given by model $M$ is $P_M$. Given a text $\boldsymbol{x}$ to be detected and its corresponding prompt $\boldsymbol{a}$, we derive the following using Bayes' rule:

$$P_w(w = 1|\boldsymbol{a}; \boldsymbol{x}) = \frac{P(\boldsymbol{x}|\boldsymbol{a}; w = 1)P_w(w = 1)}{\sum_{i=0,1} P(\boldsymbol{x}|\boldsymbol{a}; w = i)P_w(w = i)} \tag{1}$$

Dividing both the numerator and the denominator of the above equation by $P(\boldsymbol{x}|\boldsymbol{a}; w = 0)$, we obtain (1) equals:

$$\frac{A(\boldsymbol{x}, \boldsymbol{a})P_w(w = 1)}{P_w(w = 0) + A(\boldsymbol{x}, \boldsymbol{a})P_w(w = 1)} \tag{2}$$

where

$$A(\boldsymbol{x}, \boldsymbol{a}) = \frac{P(\boldsymbol{x}|\boldsymbol{a}; w = 1)}{P(\boldsymbol{x}|\boldsymbol{a}; w = 0)} \tag{3}$$

$P_w(w = 0)$ and $P_w(w = 1)$ are constants across different $\boldsymbol{x}$ as they are the probability of text being watermarked in the whole corpus. Therefore, the posterior likelihood (2) increases monotonically with $A(\boldsymbol{x}, \boldsymbol{a})$, So we can use $A(\boldsymbol{x}, \boldsymbol{a})$ to represent the likelihood of the watermark's presence. The numerator in (3) is the output probability of model $M$ with watermark injection, which is identical to $P_M(\boldsymbol{x}|\boldsymbol{a}; w = 1)$. The denominator is the conditional probability of human written texts. As large language models are pretrained on large amount of human written text, this probability can be approximated by $P_M(\boldsymbol{x}|\boldsymbol{a}; w = 0)$. (We acknowledge that this approximation may impact the performance of our detection method. Nevertheless, due to the inaccessibility of the true distribution of human-written text, we adopt this approach. Additional analysis is presented in Appendix B.) Consequently, our focus reduces to computing:

$$\hat{A}(\boldsymbol{x}, \boldsymbol{a}) = \frac{P_M(\boldsymbol{x}|\boldsymbol{a}; w = 1)}{P_M(\boldsymbol{x}|\boldsymbol{a}; w = 0)} \tag{4}$$

Applying the chain-rule $P_M(\boldsymbol{x}|\boldsymbol{a}; w) = \prod P_M(x_t|\boldsymbol{a}, x_{:t}; w)$ ($x_{:0}$ denotes empty sequence) and taking the log-transform, we have:

$$\log \hat{A}(\boldsymbol{x}, \boldsymbol{a}) = \sum_{t=0}^{T-1} \big(\log P_M(x_t|\boldsymbol{a}, x_{:t}; w = 1) - \log P_M(x_t|\boldsymbol{a}, x_{:t}; w = 0)\big) \tag{5}$$

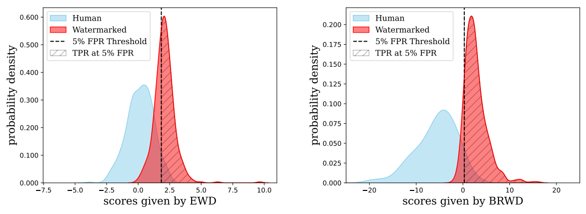
For the KGW and SWEET (assume $t$ is a high-entropy position) injection method, given the original logit of token $v$ at position $t$, $l_t^v$, the perturbed logit is $\tilde{l}_t^v = l_t^v + \delta \cdot \mathbf{1}_{\mathcal{V}_g}(v)$. For the model output distribution, we have $P_M(\cdot|\boldsymbol{a}, x_{:t}; w = 1) = \mathrm{softmax}(\tilde{\boldsymbol{l}}_t)$ and $P_M(\cdot|\boldsymbol{a}, x_{:t}; w = 0) = \mathrm{softmax}(\boldsymbol{l}_t)$. By substituting these equations of $P_M$ and $\tilde{\boldsymbol{l}}_t$ into (5), we obtain:

$$\log \hat{A}(\boldsymbol{x}, \boldsymbol{a}) = \sum_{t=0}^{T-1} \big(\delta \cdot \mathbf{1}_{\mathcal{V}_g}(x_t) - \log \frac{e^{\delta} G_t + R_t}{G_t + R_t}\big) \tag{6}$$

where $G_t = \sum_{v \in \mathcal{V}_g} e^{l_t^v}$ and $R_t = \sum_{v \in \mathcal{V}_r} e^{l_t^v}$ are the unnormalized total probability of green, red tokens, respectively.

When performing watermark detection using Equation 5, the logits $\boldsymbol{l_t}$ can be generated by a lightweight surrogate language model for efficiency. In cases where the prompt $\boldsymbol{a}$ is unavailable, a generic prompt can be used as a substitute. To clarify the detection process, we provide an algorithm description in appendix C.

Similar to the discussion on EWD, we can interpret the score of a token as reward or penalty depending on its color. Figure 2b demonstrates that the reward given by BRWD decreases monotonically, while the penalty increases monotonically with the probability of token being green, consistent with the intuition in 4.1.



(a) EWD detector (b) BRWD detector

Figure 4: Distribution of scores given by EWD and BRWD. We use MBPP dataset, StarCoder2 model and KGW watermark injection method for this experiment.

We also present the text-level score distributions of EWD and BRWD. Figures 4a and 4b show that BRWD exhibits stronger discriminative ability.

### 4.3 Theoretical Analysis

We theoretically prove that BRWD has optimal true positive rate within any false positive rate constraint, given a prompt $\boldsymbol{a}$ and model $M$. To formalize this, we need the following notations and definitions.

- Let $\Omega$ denote the set of all possible token sequence $\boldsymbol{x}$. Conditional probabilities $P_M(\cdot|\boldsymbol{a}; w = 0), P_M(\cdot|\boldsymbol{a}; w = 1)$ given by $M$ and its watermarked counterpart both endow $\Omega$ with a probability measure.

- For any watermark detection method $\phi$ which is a binary classifier on $\Omega$, its effect is equivalent to a split of $\Omega$:

  $$\Omega = \mathcal{K}_\phi \cup \mathcal{K}_\phi^c$$

  where $\mathcal{K}_\phi$ means the subset of sequences identified as watermarked by $\phi$. For our method, $\phi$=BRWD, and the detection effect is expressed by $\mathcal{K}_{\text{BRWD}}$.

**Definition 4.1.** For a detection method $\phi$, its false positive rate $\alpha_\phi$ on $\Omega(P_M(\cdot|\boldsymbol{a}))$ is:

$$\int_\Omega \mathbf{1}_{\mathcal{K}_\phi}(\boldsymbol{x}) P_M(\boldsymbol{x}|\boldsymbol{a}; w = 0)\ \mathrm{d}\boldsymbol{x}$$

**Definition 4.2.** For a detection method $\phi$, its true positive rate $\beta_\phi$ on $\Omega(P_M(\cdot|\boldsymbol{a}; w = 1))$ is:

$$\int_\Omega \mathbf{1}_{\mathcal{K}_\phi}(\boldsymbol{x}) P_M(\boldsymbol{x}|\boldsymbol{a}; w = 1)\ \mathrm{d}\boldsymbol{x}$$

**Definition 4.3.** For a threshold $\eta > 0$, the BRWD detector with this threshold, denoted as $\mathrm{BRWD}(\eta)$, classify any sequence $\boldsymbol{x}$ with score $\hat{A}(\boldsymbol{x}, \boldsymbol{a}) > \eta$ as watermarked. Therefore, its positive sample subset is:

$$\mathcal{K}_{\mathrm{BRWD}(\eta)} = \{\boldsymbol{x} \in \Omega | \hat{A}(\boldsymbol{x}, \boldsymbol{a}) > \eta\}$$

**Theorem 4.1.** *For any limit on false positive rate $\sigma > 0$, if there exists a threshold $\eta > 0$ such that $\alpha_{\mathrm{BRWD}(\eta)} = \sigma$, then for any detection method $\phi$ s.t. $\alpha_\phi \leq \sigma$, its true positive rate is no greater than that of* $\mathrm{BRWD}(\eta)$

$$\beta_\phi \leq \beta_{\mathrm{BRWD}(\eta)}$$

A proof is available in appendix D.

## 5 Experiments

We conduct experiments across different datasets, models and watermark injection methods.

**Tasks and Datasets.** We select two tasks: code generation and math problem solving. For code generation, we use HumanEval (Chen et al., 2021) and MBPP (Austin et al., 2021) datasets which contain python programming problems and reference answers. For math problem solving, we select 500 samples from GSM8K (Cobbe et al., 2021), which contains mathematical problems in English. In all experiments, we use texts longer than 15 tokens for detection.

**Models.** For code generation, we employ StarCoder2-7B (Lozhkov et al., 2024), a model specializes in programming, following prior research (Lee et al., 2024; Lu et al., 2024). For mathematical problem solving, we use DeepSeekMath-7B-Instruct (Shao et al., 2024), which is optimized for mathematical reasoning. To assess the generalizability of our method, we also utilize the versatile Llama2-7B model (Touvron et al., 2023), applying it to both code generation and math problem-solving tasks.

**Watermark injection Methods.** We utilize two injection methods, KGW and SWEET. KGW acts as a classic baseline in this field, while SWEET is specifically optimized for low-entropy scenarios. Detailed configurations are provided in Appendix E. For clarity, We use subscripts I and D to denote the corresponding injection and detection methods (e.g., $\mathrm{KGW_I}$, $\mathrm{KGW_D}$).

**Baselines and Metrics.** To detect watermarks injected by $\mathrm{KGW_I}$ and $\mathrm{SWEET_I}$, we employ their corresponding detectors as baselines. We also apply EWD as a detector for both, as it is optimized for low-entropy scenarios. Additionally, we employ the detection method proposed by DIPMark (Wu et al., 2024) as another baseline, as it does not rely on the classical z-score framework and is adaptable to other injection methods. This detector is denoted as $\mathrm{DIP_D}$. Following Lee et al. (2024) and Lu et al. (2024), we use true positive rates at 1% and 5% false positive rates(TPR@1%FPR, TPR@5%FPR), along with the corresponding F1-score, as our evaluation metrics. True positive rate (TPR) and false positive rate (FPR) represent the proportion of watermarked text successfully detected and the proportion of human-written text mistakenly classified as AI-generated, respectively. Best F1 scores are also reported to show the overall performance of detectors.

**Requirement for Detection** Our method and baselines, EWD and SWEET, are identical in that they rely on logits generated by a language model for detection. These logits can be generated using a smaller surrogate model for efficiency. Computational experiment in Append G shows nearly equivalent run time of the three detection approaches. Traditional methods like KGW detector don't need a language model for detection, but perform significantly worse than those that do in low-entropy scenarios.

## 6 Results

In this section, we present and analyze experimental results from multiple settings and perspectives.

### 6.1 Performance with Original Prompts

Table 1 and 2 (all metrics are shown in percentage) demonstrate that our method exhibits substantial advantages in detection accuracy for code and mathematical text compared to baselines, and it is applicable to different models and watermark injection methods.

For code generation task with StarCoder2 model and $\mathrm{KGW_I}$ injection method, the average improvements of BRWD over the best baseline, on two datasets are 14% in TPR@1%FPR and 24% in TPR@%5FPR. For the same task and model with $\mathrm{SWEET_I}$ injection method, the avergae improvements of the two metrics are 23% and 27%, respectively. For mathematical problem solving task with DeepSeek model, the improvements of the two metrics are 38%, 19% while using $\mathrm{KGW_I}$ injection method and 35%, 24% while using $\mathrm{SWEET_I}$ injection method. Notably, for this model, our method

| Model | Injection | Detection | HumanEval |  |  |  |  | MBPP |  |  |  |  |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
|  |  |  | 1%FPR |  | 5%FPR |  | Best | 1%FPR |  | 5%FPR |  | Best |
|  |  |  | TPR | F1 | TPR | F1 | F1 | TPR | F1 | TPR | F1 | F1 |
| StarCoder2 | $KGW_I$ | $KGW_D$ | 10.92 | 19.55 | 29.41 | 44.03 | 73.87 | 8.02 | 14.74 | 24.30 | 37.58 | 74.05 |
|  |  | $DIP_D$ | 17.65 | 29.79 | 29.41 | 44.03 | 73.22 | 19.09 | 31.83 | 35.57 | 50.62 | 73.81 |
|  |  | EWD | 44.54 | 61.27 | 60.50 | 73.47 | 84.13 | 35.14 | 51.67 | 66.38 | 77.47 | 88.28 |
|  |  | BRWD | **63.87** | **77.55** | **83.19** | **88.79** | **93.98** | **43.60** | **60.36** | **91.32** | **93.04** | **94.25** |
|  | $SWEET_I$ | $SWEET_D$ | 32.23 | 48.45 | 49.59 | 64.52 | 80.74 | 37.09 | 53.77 | 54.66 | 68.57 | 84.84 |
|  |  | $DIP_D$ | 9.09 | 16.54 | 32.23 | 47.27 | 82.96 | 35.36 | 51.91 | 67.68 | 78.39 | 86.71 |
|  |  | EWD | 38.84 | 55.62 | 54.55 | 68.39 | 83.97 | 32.75 | 49.03 | 62.69 | 74.77 | 88.21 |
|  |  | BRWD | **58.68** | **73.58** | **83.47** | **88.60** | **92.37** | **59.00** | **73.81** | **88.29** | **91.36** | **93.97** |
| Llama2 | $KGW_I$ | $KGW_D$ | 58.78 | 73.73 | 74.32 | 83.02 | 84.93 | 28.43 | 43.99 | 61.09 | 73.63 | 84.17 |
|  |  | $DIP_D$ | 21.62 | 35.36 | 62.16 | 74.49 | 80.23 | 34.68 | 51.19 | 61.29 | 73.79 | 83.01 |
|  |  | EWD | 96.62 | 97.95 | 98.65 | 97.01 | 97.95 | 50.81 | 67.02 | 79.64 | 86.34 | 91.07 |
|  |  | BRWD | **97.97** | **98.64** | **100** | **97.69** | **99.00** | **66.33** | **79.37** | **98.59** | **96.93** | **97.11** |
|  | $SWEET_I$ | $SWEET_D$ | 84.67 | 91.37 | 90.67 | 92.83 | 94.77 | 49.09 | 65.50 | 77.17 | 84.79 | 88.62 |
|  |  | $DIP_D$ | 77.37 | 86.89 | 86.67 | 90.59 | 93.46 | 60.81 | 75.25 | 88.08 | 91.31 | 90.25 |
|  |  | EWD | 85.33 | 91.76 | 90.67 | 92.83 | 94.53 | 50.10 | 66.40 | 84.04 | 88.98 | 90.84 |
|  |  | BRWD | **92.00** | **95.50** | **96.00** | **95.68** | **96.75** | **83.64** | **90.69** | **97.98** | **96.61** | **97.19** |

Table 1: Detection accuracy on HumanEval, MBPP datasets using StarCoder2-7B, Llama2-7B models and $KGW_I$, $SWEET_I$ injection methods.

| Model | Injection | Detection | GSM8K |  |  |  |  |
|---|---|---|---|---|---|---|---|
|  |  |  | 1%FPR |  | 5%FPR |  | Best |
|  |  |  | TPR | F1 | TPR | F1 | F1 |
| DeepSeek | $KGW_I$ | $KGW_D$ | 8.00 | 14.68 | 24.20 | 37.46 | 67.73 |
|  |  | $DIP_D$ | 1.80 | 3.47 | 7.40 | 13.17 | 67.71 |
|  |  | EWD | 52.60 | 68.49 | 78.20 | 85.37 | 87.86 |
|  |  | BRWD | **91.00** | **94.79** | **97.60** | **96.35** | **96.83** |
|  | $SWEET_I$ | $SWEET_D$ | 57.60 | 72.64 | 75.40 | 83.59 | 88.21 |
|  |  | $DIP_D$ | 31.20 | 47.20 | 69.40 | 79.86 | 87.23 |
|  |  | EWD | 40.40 | 57.14 | 70.40 | 80.27 | 85.47 |
|  |  | BRWD | **92.80** | **95.77** | **99.80** | **97.46** | **97.65** |
| Llama2 | $KGW_I$ | $KGW_D$ | 73.35 | 84.23 | 81.16 | 87.28 | 88.35 |
|  |  | $DIP_D$ | 59.32 | 74.09 | 71.34 | 81.00 | 82.80 |
|  |  | EWD | 90.38 | 94.55 | 94.79 | 94.98 | 95.93 |
|  |  | BRWD | **90.98** | **94.88** | **95.99** | **95.61** | **95.96** |
|  | $SWEET_I$ | $SWEET_D$ | 84.77 | 91.36 | 93.39 | 94.24 | 94.27 |
|  |  | $DIP_D$ | 88.38 | 93.43 | 94.79 | 95.36 | 94.83 |
|  |  | EWD | 89.58 | 94.11 | 95.39 | 95.30 | 95.94 |
|  |  | BRWD | **95.99** | **97.56** | **97.60** | **96.44** | **97.57** |

Table 2: Detection accuracy on GSM8K dataset using DeepSeek-math-7B-instruct, Llama2-7B models and $KGW_I$, $SWEET_I$ watermark injection methods.

improves the TPR@1%FPR from below 60% to over 90%.

For Llama2 model, our detection method also outperforms all baselines. Specifically, it achieves an improvement of 34% in TPR@1%FPR on MBPP dataset while using $SWEET_I$ injection method. Although our method shows relatively small advantages in certain settings, the baselines have already achieved high accuracy in these settings, leaving limited room for improvement.

## 6.2 Performance without original Prompts

In applications like cheating detection, prompts are usually accessible, specifically the question text. However, for code copyright protection, the original prompt is often unavailable. Therefore, we also conduct experiments with a general prompt following previous studies (Lee et al., 2024; Lu et al., 2024). The details about this setting are in Appendix E.2. Table 3 and Table 4 show that our method still outperforms all baselines. For the code generation task and $KGW_I$ watermark injection method, BRWD achieves average improvements of 8.49% and 16.28% in TPR@1%FPR and TPR@5%FPR over the best baseline, respectively. For the same task and $SWEET_I$ method, BRWD demonstrates average improvements of 6.61% and 16.45% in TPR@1%FPR and TPR@5%FPR, respectively. For mathematical task and $KGW_I$ method, BRWD achieves improvements of 15.68% and 11.81% for the two metrics, respectively. For this task with $SWEET_I$ method, BRWD results in improvements of 4.48% and 6.31% for the same metrics. We also find that as text length increases, the absence of the original prompts has a diminishing impact on detection accuracy.

## 6.3 Adaptability to High-Entropy Texts

To evaluate the applicability of our detection method in general high-entropy scenario, we con-

| Injection | Detection | HumanEval | | | | | MBPP | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 1%FPR | | 5%FPR | | Best | 1%FPR | | 5%FPR | | Best |
| | | TPR | F1 | TPR | F1 | F1 | TPR | F1 | TPR | F1 | F1 |
| $\text{KGW}_\text{I}$ | $\text{KGW}_\text{D}$ | 9.65 | 17.46 | 28.95 | 43.42 | 73.57 | 8.03 | 14.74 | 24.30 | 37.58 | 74.05 |
| | $\text{DIP}_\text{D}$ | 12.28 | 21.71 | 28.07 | 42.38 | 73.57 | 19.09 | 31.83 | 35.57 | 50.62 | 73.81 |
| | EWD | 24.56 | 39.16 | 42.11 | 57.49 | 80.65 | 17.14 | 29.04 | 36.88 | 51.99 | 77.77 |
| | BRWD | **29.82** | **45.64** | **44.74** | **60.00** | **87.10** | **28.85** | **44.48** | **66.81** | **77.78** | **86.76** |
| $\text{SWEET}_\text{I}$ | $\text{SWEET}_\text{D}$ | 19.83 | 32.86 | 37.93 | 53.33 | 77.65 | 9.33 | 16.93 | 32.54 | 47.32 | 77.32 |
| | $\text{DIP}_\text{D}$ | 6.03 | 11.29 | 33.62 | 48.75 | 77.78 | 20.17 | 33.33 | 35.57 | 51.01 | 76.25 |
| | EWD | 23.28 | 37.50 | 40.52 | 55.95 | 79.70 | 14.75 | 25.52 | 36.88 | 51.99 | 78.25 |
| | BRWD | **25.86** | **40.82** | **50.00** | **64.80** | **84.43** | **25.38** | **40.21** | **60.30** | **72.97** | **86.33** |

Table 3: Detection accuracy on HumanEval, MBPP datasets using StarCoder2-7B model and $\text{KGW}_\text{I}$, $\text{SWEET}_\text{I}$ injection methods without original prompts.

| Injection | Detection | GSM8K | | | | |
|---|---|---|---|---|---|---|
| | | 1%FPR | | 5%FPR | | Best |
| | | TPR | F1 | TPR | F1 | F1 |
| $\text{KGW}_\text{I}$ | $\text{KGW}_\text{D}$ | 7.13 | 13.21 | 23.83 | 37.03 | 67.65 |
| | $\text{DIP}_\text{D}$ | 1.8 | 3.47 | 7.40 | 13.17 | 67.71 |
| | EWD | 11.61 | 20.65 | 40.53 | 55.74 | 78.32 |
| | BRWD | **27.29** | **42.61** | **52.34** | **66.58** | **78.80** |
| $\text{SWEET}_\text{I}$ | $\text{SWEET}_\text{D}$ | 8.35 | 15.30 | 28.92 | 43.23 | 73.74 |
| | $\text{DIP}_\text{D}$ | 7.20 | 13.31 | 22.40 | 35.16 | 73.63 |
| | EWD | 7.13 | 13.21 | 30.96 | 45.58 | **75.05** |
| | BRWD | **12.83** | **22.58** | **37.27** | **52.44** | 74.14 |

Table 4: Detection accuracy on GSM8K dataset using DeepSeek-math-7B-instruct model and $\text{KGW}_\text{I}$, $\text{SWEET}_\text{I}$ injection methods without original prompts.

| Injection | Detection | 1%FPR | | 5%FPR | | Best |
|---|---|---|---|---|---|---|
| | | TPR | F1 | TPR | F1 | F1 |
| $\text{KGW}_\text{I}$ | $\text{KGW}_\text{D}$ | 99.59 | 99.39 | 99.59 | 97.42 | 99.39 |
| | $\text{DIP}_\text{D}$ | 98.99 | 99.09 | 99.59 | 97.42 | 98.59 |
| | EWD | 99.59 | 99.39 | **100** | **97.62** | 99.70 |
| | BRWD | **100** | **99.60** | **100** | **97.62** | **100** |
| $\text{SWEET}_\text{I}$ | $\text{SWEET}_\text{D}$ | 99.80 | 99.50 | 99.80 | 97.53 | 99.70 |
| | $\text{DIP}_\text{D}$ | **100** | **99.60** | **100** | **97.63** | 99.90 |
| | EWD | 99.80 | 99.50 | **100** | **97.63** | 99.70 |
| | BRWD | **100** | **99.60** | **100** | **97.63** | **100** |

Table 5: Detection performance on C4 dataset using Llama2-7B model and $\text{KGW}_\text{I}$, $\text{SWEET}_\text{I}$ watermark injection methods.

duct experiments using 500 samples from C4 (Raffel et al., 2023) English news dataset and the Llama2-7B model. Detailed settings are in Appendix E.3. Table 5 shows that our detection method outperforms the baselines.

### 6.4 Robustness against Removal Attack

Since malicious users tend to modify the content generated by LLMs to escape detection, we need to assess the robustness of our detector against removal attacks. Following Lee et al. (2024), we use variable name substitution as our attack method. We replace 50% of the variables in each function with random strings of 2 to 5 characters (Results under more intense attack setting are provided in Appendix H). The watermark injection method is $\mathbf{KGW_I}$ with $\delta = 2$ and $\gamma = 0.5$. All methods show a significant decrease in accuracy after the attack, but our method remains the best overall.

| Detection | 1%FPR | | 5%FPR | | Best |
|---|---|---|---|---|---|
| | TPR | F1 | TPR | F1 | F1 |
| $\text{KGW}_\text{D}$ | 5.51 | 10.35 | 13.62 | 22.98 | 70.60 |
| $\text{DIP}_\text{D}$ | 10.43 | 18.75 | 15.94 | 26.38 | 68.84 |
| EWD | **15.94** | **27.30** | 29.86 | 44.30 | 76.38 |
| BRWD | 14.78 | 25.56 | **46.96** | **61.83** | **79.83** |

Table 6: Detection performance on MBPP dataset after variable substitution attack. The model and watermark injection method used are StarCoder2-7B and $\text{KGW}_\text{I}$

## 7 Conclusion

In this study, we identify the misalignment in entropy-based watermark detectors and propose a detector that addresses this problem. Our detector utilizes Bayes' rule to fully leverage the distribution given by language model. We provide theoretical analysis on its optimality and empirically verify its superiority in low-entropy scenario. Across various models and watermark injection methods, under scenarios with or without prompts, our method demonstrates superior accuracy and robustness. Notably, compared to the best baseline, our method achieves up to 1.5 times accuracy on the code datasets and up to 1.7 times accuracy on the mathematics dataset.

## 8 Limitations

Our detection method has two main limitations. First, the generation tasks and datasets tested are limited. We employ two code datasets and one mathematics dataset for evaluation. We plan to test our method on a broader range of low-entropy tasks with more diverse datasets. Second, we only test two watermark injection methods in this study. Our approach is applicable to a variety of generative watermark injection methods based on distribution perturbation. We plan to implement and evaluate our method with other watermark injection techniques in future work.

## 9 Acknowledgement

This work was supported by the Strategic Priority Research Program of the CAS under Grants No. XDB0680302, the National Natural Science Foundation of China under Grant No. 62472409, and the Beijing Natural Science Foundation under Grant No. 4252023.

## References

Sahar Abdelnabi and Mario Fritz. 2021. Adversarial watermarking transformer: Towards tracing text provenance with data hiding. *Preprint*, arXiv:2009.03015.

R. Michael Alvarez, Frederick Eberhardt, and Mitchell Linegar. 2023. Generative ai and the future of elections.

Isabelle Augenstein, Timothy Baldwin, Meeyoung Cha, Tanmoy Chakraborty, Giovanni Luca Ciampaglia, David Corney, Renee DiResta, Emilio Ferrara, Scott Hale, Alon Halevy, Eduard Hovy, Heng Ji, Filippo Menczer, Ruben Miguez, Preslav Nakov, Dietram Scheufele, Shivam Sharma, and Giovanni Zagni. 2023. Factuality challenges in the era of large language models. *Preprint*, arXiv:2310.05189.

Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie Cai, Michael Terry, Quoc Le, and Charles Sutton. 2021. Program synthesis with large language models. *Preprint*, arXiv:2108.07732.

Damian Canagasuriam and Eden-Raye Lukacik. 2024. Chatgpt, can you take my job interview? examining artificial intelligence cheating in the asynchronous video interview. *International Journal of Selection and Assessment*, 33(1):e12491.

Liang Chen, Yatao Bian, Yang Deng, Deng Cai, Shuaiyi Li, Peilin Zhao, and Kam-Fai Wong. 2024. WatME: Towards lossless watermarking through lexical redundancy. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9166–9180, Bangkok, Thailand. Association for Computational Linguistics.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew N. Carr, Jan Leike, Josh Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. 2021. Evaluating large language models trained on code. *Preprint*, arXiv:2107.03374.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *Preprint*, arXiv:2110.14168.

Sumanth Dathathri, Abigail See, Sumedh Ghaisas, Po-Sen Huang, Rob McAdam, Johannes Welbl, Vandana Bachani, Alex Kaskasoli, Robert Stanforth, Tatiana Matejovicova, Jamie Hayes, Nidhi Vyas, Majd Al Merey, Jonah Brown-Cohen, Rudy Bunel, Borja Balle, Taylan Cemgil, Zahra Ahmed, Kitty Stacpoole, Ilia Shumailov, Ciprian Baetu, Sven Gowal, Demis Hassabis, and Pushmeet Kohli. 2024. Scalable watermarking for identifying large language model outputs. *Nature*, 634:818–823.

Pierre Fernandez, Antoine Chaffin, Karim Tit, Vivien Chappelier, and Teddy Furon. 2023. Three bricks to consolidate watermarks for large language models. *2023 IEEE International Workshop on Information Forensics and Security (WIFS)*.

Daniel Fried, Armen Aghajanyan, Jessy Lin, Sida Wang, Eric Wallace, Freda Shi, Ruiqi Zhong, Wen tau Yih, Luke Zettlemoyer, and Mike Lewis. 2023. Incoder: A generative model for code infilling and synthesis. *Preprint*, arXiv:2204.05999.

Mingjia Huo, Sai Ashish Somayajula, Youwei Liang, Ruisi Zhang, Farinaz Koushanfar, and Pengtao Xie. 2024. Token-specific watermarking with enhanced detectability and semantic coherence for large language models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 20746–20767. PMLR.

John Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, and Tom Goldstein. 2023. A watermark for large language models. *The Fortieth International Conference on Machine Learning*.

Rohith Kuditipudi, John Thickstun, Tatsunori Hashimoto, and Percy Liang. 2023. Robust distortion-free watermarks for language models. *Trans. Mach. Learn. Res.*, 2024.

Taehyun Lee, Seokhee Hong, Jaewoo Ahn, Ilgee Hong, Hwaran Lee, Sangdoo Yun, Jamin Shin, and Gunhee Kim. 2024. Who wrote this code? watermarking for code generation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4890–4911, Bangkok, Thailand. Association for Computational Linguistics.

Aiwei Liu, Leyi Pan, Xuming Hu, Shiao Meng, and Lijie Wen. 2024a. A semantic invariant robust watermark for large language models. *Preprint*, arXiv:2310.06356.

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

Yepeng Liu and Yuheng Bu. 2024. Adaptive text watermark for large language models. *Preprint*, arXiv:2401.13927.

Anton Lozhkov, Raymond Li, Loubna Ben Allal, Federico Cassano, Joel Lamy-Poirier, Nouamane Tazi, Ao Tang, Dmytro Pykhtar, Jiawei Liu, Yuxiang Wei, Tianyang Liu, Max Tian, Denis Kocetkov, Arthur Zucker, Younes Belkada, Zijian Wang, Qian Liu, Dmitry Abulkhanov, Indraneil Paul, Zhuang Li, Wen-Ding Li, Megan Risdal, Jia Li, Jian Zhu, Terry Yue Zhuo, Evgenii Zheltonozhskii, Nii Osae Osae Dade, Wenhao Yu, Lucas Krauß, Naman Jain, Yixuan Su, Xuanli He, Manan Dey, Edoardo Abati, Yekun Chai, Niklas Muennighoff, Xiangru Tang, Muhtasham Oblokulov, Christopher Akiki, Marc Marone, Chenghao Mou, Mayank Mishra, Alex Gu, Binyuan Hui, Tri Dao, Armel Zebaze, Olivier Dehaene, Nicolas Patry, Canwen Xu, Julian McAuley, Han Hu, Torsten Scholak, Sebastien Paquet, Jennifer Robinson, Carolyn Jane Anderson, Nicolas Chapados, Mostofa Patwary, Nima Tajbakhsh, Yacine Jernite, Carlos Muñoz Ferrandis, Lingming Zhang, Sean Hughes, Thomas Wolf, Arjun Guha, Leandro von Werra, and Harm de Vries. 2024. Starcoder 2 and the stack v2: The next generation. *Preprint*, arXiv:2402.19173.

Yijian Lu, Aiwei Liu, Dianzhi Yu, Jingjing Li, and Irwin King. 2024. 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, Bangkok, Thailand. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2023. Exploring the limits of transfer learning with a unified text-to-text transformer. *Preprint*, arXiv:1910.10683.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *Preprint*, arXiv:2402.03300.

Teo Susnjak. 2022. Chatgpt: The end of online exam integrity? *Preprint*, arXiv:2212.09292.

Umut Topkara, Mercan Topkara, and Mikhail J. Atallah. 2006. The hiding virtues of ambiguity: quantifiably resilient watermarking of natural language text through synonym substitutions. In *Proceedings of the 8th Workshop on Multimedia and Security*, MM and Sec '06, page 164–174, New York, NY, USA. Association for Computing Machinery.

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

Lean Wang, Wenkai Yang, Deli Chen, Hao Zhou, Yankai Lin, Fandong Meng, Jie Zhou, and Xu Sun. 2024. Towards codable watermarking for injecting multi-bits information to llms. *Preprint*, arXiv:2307.15992.

Yihan Wu, Zhengmian Hu, Junfeng Guo, Hongyang Zhang, and Heng Huang. 2024. A resilient and accessible distribution-preserving watermark for large language models.

Xi Yang, Kejiang Chen, Weiming Zhang, Chang Liu, Yuang Qi, Jie Zhang, Han Fang, and Nenghai Yu. 2023. Watermarking text generated by black-box language models. *arXiv preprint arXiv:2305.08883*.

KiYoon Yoo, Wonhyuk Ahn, Jiho Jang, and Nojun Kwak. 2023. Robust multi-bit natural language watermarking through invariant features. *Preprint*, arXiv:2305.01904.

KiYoon Yoo, Wonhyuk Ahn, and Nojun Kwak. 2024. Advancing beyond identification: Multi-bit watermark for large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4031–4055, Mexico City, Mexico. Association for Computational Linguistics.

Ruisi Zhang, Shehzeen Samarah Hussain, Paarth Neekhara, and Farinaz Koushanfar. 2024. Remark-llm: A robust and efficient watermarking framework for generative large language models. In *33rd USENIX Security Symposium (USENIX Security 24)*, pages 1813–1830.

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

## A Notations

| Symbol | Description |
|---|---|
| $M$ | Language model used for generation. |
| $\mathcal{V}$ | Vocabulary of $M$. |
| $\mathcal{V}_g$ | Green list. |
| $\mathcal{V}_r$ | Red list. |
| $v$ | Token in $V$. |
| $w$ | Whether watermark exists. (0/1) |
| $P_M$ | Distribution given by $M$. |
| $P$ | Distribution of all natural language. |
| $P_w$ | Distribution of $w$. |
| $\boldsymbol{a}$ | Prompt used for generation. |
| $\boldsymbol{x}$ | Text generated by $M$. |
| $x_t$ | The $t$-th token in $\boldsymbol{x}$. |
| $t$ | Generation step. |
| $\boldsymbol{l}_t$ | Model logits at the $t$-th step. |
| $l_t^v$ | Model logit of token $v$ at the $t$-th step. |
| $\boldsymbol{p}_t$ | Distribution given by $M$ at step $t$. |
| $p_t^v$ | Probability of $v$ given by $M$ at step $t$. |
| $H_t$ | Information entropy at step $t$. |
| $S(\boldsymbol{x})$ | Watermark detection score of $\boldsymbol{x}$. |
| $s(\boldsymbol{x}, t)$ | Watermark detection score given to $x_t$. |
| $\mathrm{norm}$ | Normalizer for watermark detection. |
| $T$ | Length of $\boldsymbol{x}$. |
| $\gamma$ | Green-list ratio in KGW and SWEET. |
| $\delta$ | $\delta$ in KGW and SWEET. |

Table 7: Notations used throughout the paper.

## B Analysis of Approximation

Our additional experiment examines the impact of the approximation of human text distribution, under varying levels of divergence between the true distribution and the estimated distribution.

In this experiment, the key challenge is to acquire the true distribution of unwatermarked text, which is unattainable in practice. Therefore, we utilized a simulated unwatermarked corpus, allowing us to precisely control and analyze the true distribution.

We generate the simulated unwatermarked corpora using Llama-2-7B and the prompts in the HumanEval dataset, introducing a controllable additive perturbation to its output logits. The perturbation follows a uniform distribution $U[0, s]$, where $s$ serves as a variable parameter to regulate the level of divergence. We denote the output distributions of perturbed models as $P_t$, which is the true distribution of unwatermarked text. And the estimated distribution of unwatermarked text is $P_e$, which is the the original distribution of Llama-2-7B model.

To measure the divergence between true distribution $P_t$ and estimated distributions $P_e$, we compute the average perplexity of unwatermarked samples in the simulated corpus using the Llama-2-7B model. We also compute the average perplexity of human-written completions in HumanEval dataset. As perplexity indicates the extent to which a text diverges from the model's distribution, a synthetic corpus exhibiting a perplexity close to that of human-written corpus is likely to incur a similar level of error when $P_e$ is used to approximate $P_t$ in our method.

To quantify the effect of distribution approximation, we calculated the rigorous watermark detection score $\log A(\boldsymbol{x}, \boldsymbol{a})$ (computed using $P_t$) and approximated detection score $\log \hat{A}(\boldsymbol{x}, \boldsymbol{a})$ (computed using $P_e$) for both watermarked and unwatermarked texts. Then we can compare two sets of scores to analyze the effect of distribution approximation.

As shown in Table 8, regardless of whether $\log \hat{A}$ or $\log A$ is used, the scores exhibit a clear distinction between unwatermarked samples and watermarked samples.

Meanwhile, when the perplexity is similar to human corpus's (1.62), there is an observable difference in the score gap of watermarked and unwatermarked text, i.e. $\log A(\text{watermarked}) - \log A(\text{unwatermarked})$ is larger than $\log \hat{A}(\text{watermarked}) - \log \hat{A}(\text{unwatermarked})$. For instance, when noise scale $s = 5.75$, the mean gap of $\log \hat{A}(\text{watermarked})$ and $\log \hat{A}(\text{unwatermarked})$ is 16.46 while it is 89.24 for the rigorous scores, showing that detection performance can be further improved by using more accurate approximation.

## C Algorithm

We present the BRWD detection procedure for the KGW watermark injection method in Algorithm 1. For the SWEET method, it suffices to incorporate a low-entropy token filter inside the for-loop.

## D Proof of Theorem

Theorem 4.1 means that for any detection method $\phi$, when its false positive rate (FPR) is lower than that of BRWD (i.e., $\alpha_\phi \le \alpha_{\mathrm{BRWD}(\eta)}$), its true positive rate (TPR) is also lower than that of BRWD (i.e., $\beta_\phi \le \beta_{\mathrm{BRWD}(\eta)}$).

To prove Theorem 4.1, it suffices to prove:

$$\beta_{\mathrm{BRWD}(\eta)} - \beta_\phi \ge \eta(\alpha_{\mathrm{BRWD}(\eta)} - \alpha_\phi)$$

| noise scale | PPL | source | detection | mean | 1/4 quantile | 1/2 quantile | 3/4 quantile |
|---|---|---|---|---|---|---|---|
| 5.5 | 1.52 | watermarked | $\log \hat{A}$ | 9.15 | 3.56 | 7.18 | 11.25 |
| | | | $\log A$ | 54.01 | 18.11 | 34.56 | 58.01 |
| | | unwatermarked | $\log \hat{A}$ | -8.42 | -15.08 | -9.89 | -1.45 |
| | | | $\log A$ | -28.53 | -41.88 | -31.09 | -14.26 |
| 5.75 | 1.57 | watermarked | $\log \hat{A}$ | 9.15 | 3.56 | 7.18 | 11.25 |
| | | | $\log A$ | 58.35 | 19.59 | 37.43 | 62.79 |
| | | unwatermarked | $\log \hat{A}$ | -7.31 | -15.77 | -7.29 | 1.59 |
| | | | $\log A$ | -30.89 | -46.96 | -28.60 | -9.43 |
| 6.0 | 1.67 | watermarked | $\log \hat{A}$ | 9.15 | 3.56 | 7.18 | 11.25 |
| | | | $\log A$ | 62.87 | 21.15 | 40.70 | 67.26 |
| | | unwatermarked | $\log \hat{A}$ | -5.54 | -12.33 | -5.55 | 3.01 |
| | | | $\log A$ | -28.23 | -43.14 | -30.18 | -6.45 |

Table 8: Approximation analysis on HumanEval dataset and simulated unwatermarked corpus.

**Algorithm 1** BRWD Detection on KGW

```
Require: Language model $M$, target text $\boldsymbol{x}$, detection key $k$, window size $m$, detection threshold $\tau$, green list ratio $\gamma$, watermark bias $\delta$
Ensure: 1 if input text is watermarked, else 0
1: Compute distribution $P_M = [p_0, p_1, \dots]$ for each token using $\boldsymbol{x}$ and $M$.
2: $s \leftarrow 0$
3: for $l = m, m+1, \dots$ do
4:     Use $k$ and previous $m$ tokens $x_{l-m:l-1}$ to find green list $\mathcal{V}_g$ and red list $\mathcal{V}_r$
5:     Get green/red list total probability $P(\mathcal{V}_g)$ and $P(\mathcal{V}_r)$ using $P_M$
6:     $s \leftarrow s - \log\left(e^{\delta} P(V_G) + P(V_R)\right)$
7:     if $x_l \in \mathcal{V}_g$ then
8:         $s \leftarrow s + \delta$
9:     end if
10: end for
11: if $s > \tau$ then
12:     return 1
13: else
14:     return 0
15: end if
```

From Def 4.1 and 4.2, the above inequality can be expressed as:

$$\int_{\Omega} \mathbf{1}_{\mathcal{K}_{BRWD(\eta)}}(\boldsymbol{x}) P_M(\boldsymbol{x}|\boldsymbol{a}, w=1) d\boldsymbol{x} - \int_{\Omega} \mathbf{1}_{\mathcal{K}_{\phi}}(\boldsymbol{x}) P_M(\boldsymbol{x}|\boldsymbol{a}, w=1) d\boldsymbol{x} - \eta\Big(\int_{\Omega} \mathbf{1}_{\mathcal{K}_{BRWD(\eta)}}(\boldsymbol{x}) P_M(\boldsymbol{x}|\boldsymbol{a}, w=0) d\boldsymbol{x} - \int_{\Omega} \mathbf{1}_{\mathcal{K}_{\phi}}(\boldsymbol{x}) P_M(\boldsymbol{x}|\boldsymbol{a}, w=0) d\boldsymbol{x}\Big) \geq 0$$

Using the linearity of integration and combining terms, our goal reduces to proving: $\int_{\Omega} S(\boldsymbol{x}) d\boldsymbol{x} \geq 0$, where

$$S(\boldsymbol{x}) = [\mathbf{1}_{\mathcal{K}_{\mathrm{BRWD}(\eta)}}(\boldsymbol{x}) - \mathbf{1}_{\mathcal{K}_{\phi}}(\boldsymbol{x})] \cdot [P_M(\boldsymbol{x}|\boldsymbol{a}, w=1) - \eta P_M(\boldsymbol{x}|\boldsymbol{a}, w=0)]$$

To prove this inequality, we simply classify the discussion based on whether $\boldsymbol{x}$ is in $\mathcal{K}_{\mathrm{BRWD}(\eta)}$.

- For $\boldsymbol{x} \in \mathcal{K}_{\mathrm{BRWD}(\eta)}$, we have $\mathbf{1}_{\mathcal{K}_{\mathrm{BRWD}(\eta)}}(\boldsymbol{x}) = 1$ and $\mathbf{1}_{\mathcal{K}_{\phi}}(\boldsymbol{x}) \leq 1$ from the definition of indicator function, therefore

$$\mathbf{1}_{\mathcal{K}_{\mathrm{BRWD}(\eta)}}(\boldsymbol{x}) - \mathbf{1}_{\mathcal{K}_{\phi}}(\boldsymbol{x}) \geq 0$$

From Def 4.3, we have

$$\hat{A}(\boldsymbol{x}, \boldsymbol{a}) = \frac{P_M(\boldsymbol{x}|\boldsymbol{a}, w=1)}{P_M(\boldsymbol{x}|\boldsymbol{a}, w=0)} > \eta$$

and this leads to

$$P_M(\boldsymbol{x}|\boldsymbol{a}, w=1) - \eta P_M(\boldsymbol{x}|\boldsymbol{a}, w=0) \geq 0$$

Thus $S(\boldsymbol{x}) \geq 0$.

- For $\boldsymbol{x} \notin \mathcal{K}_{\mathrm{BRWD}(\eta)}$, we can derive

  $$\mathbf{1}_{\mathcal{K}_{\mathrm{BRWD}(\eta)}}(\boldsymbol{x}) - \mathbf{1}_{\mathcal{K}_{\phi}}(\boldsymbol{x}) \geq 0$$

  and

  $$P_M(\boldsymbol{x}|\boldsymbol{a}, w = 1) - \eta P_M(\boldsymbol{x}|\boldsymbol{a}, w = 0) \geq 0$$

  in the same way. Therefore we have $S(\boldsymbol{x}) \geq 0$

Since $S(\boldsymbol{x})$ is non-negative on every point $\boldsymbol{x}$ in $\Omega$, its integral on $\Omega$ is also non-negative. Thus the proof is completed.

## E Detailed Experiment Configurations

### E.1 Generation Settings

The sampling strategy adopted is top-p sampling with top-p=0.95. We set temperature=0.2 during generation. For the HumanEval dataset, we feed the original prompts in it to language models. For the MBPP dataset, we use three-shot prompts following Fried et al. (2023). For the GSM8K dataset, we add a chain-of-thought instruction to each question to construct our prompt. The instruction we use is "Please reason step by step, and put your final answer within \boxed{}.". Regarding the watermark injecting methods, we set $\gamma = 0.5$ and $\delta = 2$ for $\mathrm{KGW_I}$ and $\mathrm{SWEET_I}$. As for the entropy-threshold in $\mathrm{SWEET_I}$, we set it to 0.65 for experiments with Llama2 model and 0.6 otherwise.

### E.2 General Prompts

For Python function completion tasks in HumanEval dataset, we used the general prompt "def solution(*args):\n "'Generate a solution"'\n". For programming tasks in MBPP dataset, we used the general prompt "Write a python function to implement a specific requirement.\n". For math problem solving tasks in GSM8K dataset, the general prompt we used is "Solve a math problem, please think by step.\n".

### E.3 High-Entropy Dataset

For both $\mathrm{KGW_I}$ and $\mathrm{SWEET_I}$, we set $\gamma = 0.5$ and $\delta = 2$. For SWEET injection, we set the entropy threshold at 0.65. The temperature is 0.7, and the sampling strategy is top-p sampling with top-p=0.95. The generated texts are truncated to 200 tokens, and the minimum length for detection is 15 tokens. The original prompts are used for detection in this experiment.

| metric → | TPR@1%FPR / TPR@5%FPR | | |
|---|---|---|---|
| parameters ↓ | $\mathrm{KGW_D}$ | EWD | BRWD |
| $\delta = 1.5, \gamma = 0.25$ | 8.5 / 13.8 | 16.9 / 46.9 | **34.6 / 56.9** |
| $\delta = 1.5, \gamma = 0.5$ | 6.5 / 22.1 | 30.3 / 42.6 | **47.5 / 63.9** |
| $\delta = 1.5, \gamma = 0.75$ | 18.7 / 22.4 | 26.2 / 33.6 | **45.8 / 56.1** |
| $\delta = 2.0, \gamma = 0.25$ | 19.4 / 31.6 | 32.3 / 59.7 | **61.9 / 83.5** |
| $\delta = 2.0, \gamma = 0.5$ | 10.9 / 29.4 | 44.5 / 60.5 | **63.9 / 83.2** |
| $\delta = 2.0, \gamma = 0.75$ | 21.6 / 31.0 | 43.9 / 68.9 | **73.3 / 79.3** |
| $\delta = 2.5, \gamma = 0.25$ | 32.1 / 48.6 | 53.6 / 85.0 | **89.3 / 95.7** |
| $\delta = 2.5, \gamma = 0.5$ | 19.1 / 47.8 | 60.8 / 75.6 | **81.7 / 94.8** |
| $\delta = 2.5, \gamma = 0.75$ | 28.5 / 38.4 | 62.5 / 76.8 | **74.1 / 91.1** |

Table 9: Watermark detection accuracy with respect to StarCoder2-7B model, KGW watermark injection method, and HumanEval dataset. All metrics presented in this table are percentages.

| Method | KGW | DIP | EWD | SWEET | BRWD |
|---|---|---|---|---|---|
| Time | 0.114 | 0.113 | 0.425 | 0.417 | 0.424 |

Table 10: This table shows the average time (in seconds) taken to detect 200-token text that is watermarked by KGW method. These results represent the average of five runs.

## F Different Watermarking Hyper Parameters

Hyper parameters, especially $\gamma$ and $\delta$, paly an important role in KGW-based watermarking methods. To verify BRWD's adaptability to different $\delta$ and $\gamma$, we conducted additional experiments using the StarCoder2-7B model, the HumanEval dataset, and the KGW watermarking method under multiple hyper parameter combinations. The results of these experiments, as shown in Table 9, demonstrate that our method consistently outperforms the baselines across various hyper parameter settings.

## G Computational Experiment

To conduct the computational experiment, we used a single NVIDIA Tesla V100 GPU, with 500 news articles from C4 dataset. For each article, the first 30 tokens were used as a prompt to generate 200 tokens. We employed the Llama2-7B model and the KGW watermarking method. The average detection time per sample is shown in Table 10.

## H Robustness Results under Extreme Attack

We conducted an experiment involving 100% variable name substitution. The other settings are the same as Table 6 in our paper. The results in Table 11 indicate that BRWD's performance is better than EWD's at 5% FPR. And at 1% FPR, BRWD's performance is comparable to EWD's. Due to the limitations of existing automated tools for other attack methods—such as the CodePal API, which frequently introduces critical errors, we exclude them from our experiments.

| **Detection** | 1%FPR | | 5%FPR | | Best |
|---|---|---|---|---|---|
| | TPR | F1 | TPR | F1 | F1 |
| $\text{KGW}_\text{D}$ | 3.18 | 6.1 | 10.69 | 18.5 | 67.5 |
| $\text{DIP}_\text{D}$ | 4.37 | 8.7 | 12.7 | 20.4 | 64.8 |
| EWD | **7.2** | **13.3** | 17.63 | 28.7 | **69.2** |
| BRWD | 6.7 | 12.4 | **23.7** | **36.9** | **69.2** |

Table 11: Detection performance on MBPP dataset after 100%variable substitution attack. The model and watermark injection method used are StarCoder2-7B and $\text{KGW}_\text{I}$