Debiasing Watermarks for Large Language Models via Maximal Coupling

Yangxinyu Xie
Department of Statistics and Data Science
University of Pennsylvania
and
Xiang Li
Department of Biostatistics, Epidemiology and Informatics
University of Pennsylvania
and
Tanwi Mallick
Mathematics and Computer Science Division
Argonne National Laboratory
and
Weijie Su
Department of Statistics and Data Science
University of Pennsylvania
and
Ruixun Zhang
School of Mathematical Sciences
Peking University

June 7, 2025

SUPPLEMENTARY MATERIAL

1 Appendix

1.1 Unbiased Watermarking Schemes and Related Works

Kirchenbauer et al. [2023a] introduced a "soft" watermark to mitigate the bias of the "hard" watermark by sampling from both the green list $\mathcal{G}$ and the red list $\mathcal{R} := \mathcal{W} \setminus \mathcal{G}$, but skewing the sampling distribution to favor tokens from the green list. That is, given a constant $\delta > 0$ and the next token distribution P_w, we construct the alternative distribution Q_w as

$$Q_w = \begin{cases} \frac{\mathrm{e}^\delta P_w}{C} & \text{if } w \in \mathcal{G} \\ \frac{P_w}{C} & \text{otherwise} \end{cases} \tag{1}$$

where $C = \mathrm{e}^\delta P_{\mathcal{G}} + P_{\mathcal{W}\setminus\mathcal{G}} = 1 + (\mathrm{e}^\delta - 1)P_{\mathcal{G}}$ is the normalizing constant. Clearly, if the green list is pre-determined, then both the "hard" and the "soft" watermarks are skewing the original token distribution. We argue that even if the green list is generated uniformly at random for each t, the soft watermark is still biased. To see that the soft watermark can distort the text distribution, consider the following example: suppose we have a binary vocabulary (that is, we only generate tokens from {0,1}) and the original language model is a biased coin, which turns up head with probability 0.9. Let the green list $\mathcal{G}$ be sampled from $\{\{0\}, \{1\}\}$ with equal probability. After embedding the soft watermark with $\delta = 1$, the sampling probability of the next token becomes

$$\begin{aligned} \mathbb{P}_{\mathcal{G}}[\text{next token is 1}] &= \mathbb{P}[\text{next token is 1}|\mathcal{G} = \{0\}]\mathbb{P}[\mathcal{G} = \{0\}] \\ &\quad + \mathbb{P}[\text{next token is 1}|\mathcal{G} = \{1\}]\mathbb{P}[\mathcal{G} = \{1\}] \\ &= \frac{1}{2} \cdot \frac{0.9}{0.9 + \mathrm{e} \cdot 0.1} + \frac{1}{2} \cdot \frac{\mathrm{e} \cdot 0.9}{\mathrm{e} \cdot 0.9 + 0.1} \approx 0.8644 \end{aligned}$$

This leads to a delicate trade-off between the text distortion and the detection power for this family of watermarking schemes, leaving the choice of δ a parameter that requires careful tuning in practice. For more discussion on such trade-off, we refer the reader to [Kirchenbauer et al., 2023a,b, Cai et al., 2024]. Our method, on the other hand, is unbiased by construction, thus avoiding this parameter-tuning issue.

Several alternative methods have been proposed to design provably unbiased watermarking schemes. The main idea involves modifying the decoding strategy of the language model decoder to create watermark signals without altering the token distributions. The first and widely recognized method adapts the exponential minimal sampling, as introduced by Aaronson and Kirchner [2023]. This method selects the next token w_t directly by computing $w_t = \mathcal{S}(\boldsymbol{P}_t, \zeta_t) := \text{argmax}_{w \in \mathcal{W}} \log U_w / P_w$ where $\zeta_t = (U_1, \ldots, U_{|\mathcal{W}|})$ is the random variable that consists of $|\mathcal{W}|$ i.i.d. copies of $U[0,1]$ and is part of the watermark key. This method is unbiased when true randomness is used for generating ζ_t and the proof involves the Gumbel trick [Gumbel, 1948, Fernandez et al., 2023, Li et al., 2024]. However, a common practice to share the watermark key is to seed a pseudorandom number generator for U with the prior k tokens $w_{(t-k)}, ..., w_{(t-1)}$ (discussed in Section **??**). Notice that upon fixing the U, the decoder is deterministic, unlike the standard decoding methods we use in this paper. When the same k-gram appears multiple times in the generated text, this deterministic decoder will generate the same token following these k-grams, leading to repetitive texts and degraded text quality [Christ et al., 2024, Kuditipudi et al., 2023]. To mitigate this, Wu et al. [2023] propose to apple repeated context masking, which prevents the watermark from being applied on step t if the same k-gram has been used to watermark previously. This method has later been adapted by Hu et al. [2023], Dathathri et al. [2024]. However, the deterministic nature of this decoder can bias the generation toward repetitive patterns

that persist even with repeated context masking, as we demonstrate in Section **??**. Several theoretical analyses and variants of this method have been proposed [Fernandez et al., 2023, Zhao et al., 2024], but no existing work completely addresses this issue.

Another method is the inverse transform sampling strategy [Christ et al., 2024, Kuditipudi et al., 2023, Hu et al., 2023, Li et al., 2024]. The idea is to first sample a standard uniform random variable $\zeta_{1:n} = (\zeta_1, ..., \zeta_n)$, and then map ζ_t to the next token. Although the randomness from ζ_t can guarantee the unbiasedness of the token generation, designing an effective test to detect watermarks with this method is challenging: Kuditipudi et al. [2023] showed that their correlation-based detection method yields weaker power than the exponential minimal sampling method, even when perfect randomness is applied, while the detection method proposed by Hu et al. [2023] involves solving an optimization problem and requires hyperparameter tuning to achieve competitive power.

Building on inverse transform sampling, Wu et al. [2023] proposed the following watermarking scheme: before generating the next token, start with a random permutation of the token set $\mathcal{W}$ and assign the last $0 < \gamma < 1$ fraction of the permutation to the green list $\mathcal{G}$. Then, reweight the token distribution P_w with a parameter $0 < \alpha < 1$ by the following strategy: first, for $i = 1, ..., |\mathcal{W}|$, define $F_{w^{(i)}} = \max\{\sum_{j=1}^{i} P_{w^{(j)}} - \alpha, 0\} + \max\{\sum_{j=1}^{i} P_{w^{(j)}} - (1-\alpha), 0\}$, where $w^{(i)}$ is the ith token in this permutation and $f(w^{(0)}) = 0$; then set $Q_{w^{(i)}} = F_{w^{(i)}} - F_{w^{(i-1)}}$. Intuitively, this token distribution $\boldsymbol{Q}$ gives more weight to tokens towards the end of the permutation. This ensures unbiased token selection after integrating the randomness of the permutation and intuitively increases the likelihood of choosing tokens from the green list. During the detection, the detector can count the number of green tokens in the generated text and compare it with γ times the total length of the text. Nonetheless, the reweighting parameter α is independent of the size γ of the green list, meaning that

reweighting doesn't necessarily favor all the green list tokens during generation, leading to a potentially noisy watermark signal. As a result, tuning these two hyperparameters, namely α and γ, becomes a delicate task.

Notably, Kuditipudi et al. [2023] advocates for the language model provider to generate the random variables $\zeta_{1:n} = (\zeta_1, ..., \zeta_n)$ before text generation and share them with the detector as part of the watermark key. The language model provider can then reuse this fixed $\zeta_{1:n}$ in clever ways to generate texts in variable lengths. When the generated text has been modified via substitution, insertion, or deletion, the detector can detect the watermark by aligning the received text with the shared random variables $\zeta_{1:n}$. We do not pursue this strategy in this paper, however, as it is not clear how the issue of multiple testing will arise from this reuse, and if so, how much it may weaken the power of the test. Moreover, aligning the received text with the shared random variables ζ may be computationally expensive for practical deployment, especially when the text is long.

Recently, several theoretical works have emerged on other aspects of watermarking for language models. Huang et al. [2023] characterizes the Uniformly Most Powerful (UMP) watermark when a small amount of distortion is allowed, as well as the minimax Type II error in the model-agnostic setting. Nonetheless, the computational efficiency of these characterized tests remains unclear. Zhang et al. [2023] demonstrates that under some assumptions, a computationally bounded attacker can erase the watermark without causing significant quality degradation. However, in many practical situations, like preventing AI-generated content from being used for further model training, existing watermarking schemes can still be useful.

1.2 Total Variation Distance and Hellinger Distance

Here we provide a brief review of the total variation distance and the Hellinger distance. For two probability measures $\boldsymbol{P}$ and $\boldsymbol{Q}$ on the same probability space $\mathcal{X}$, with densities p and q with respect to some underlying base measure ν. The total variation distance is defined as

$$\|\boldsymbol{P}-\boldsymbol{Q}\|_{\mathrm{TV}}=\frac{1}{2}\int_{\mathcal{X}}|p(x)-q(x)|d\nu(x).$$

The Hellinger distance is defined as

$$\|\boldsymbol{P}-\boldsymbol{Q}\|_{\mathrm{H}}=\left(\int_{\mathcal{X}}\left(\sqrt{p(x)}-\sqrt{q(x)}\right)^2 d\nu(x)\right)^{1/2}$$

Notice that from the definitions,

$$\frac{1}{2}\|\boldsymbol{P}-\boldsymbol{Q}\|_{\mathrm{H}}^2\leq\|\boldsymbol{P}-\boldsymbol{Q}\|_{\mathrm{TV}}.$$

and the Le Cam's inequality states that

$$\|\boldsymbol{P}-\boldsymbol{Q}\|_{\mathrm{TV}}\leq H(\boldsymbol{P}\|\boldsymbol{Q})\sqrt{1-\frac{H^2(\boldsymbol{P}\|\boldsymbol{Q})}{4}}$$

For product measures with i. i. d. components, $\boldsymbol{P}^{1:n}=\otimes_{t=1}^n\boldsymbol{P}^t$ and $\boldsymbol{Q}^{1:n}=\otimes_{t=1}^n\boldsymbol{Q}^t$, we can decouple the Hellinger distance as follows:

$$\|\boldsymbol{P}^{1:n}-\boldsymbol{Q}^{1:n}\|_{\mathrm{H}}^2=2-2\left(1-\frac{1}{2}H^2\left(\boldsymbol{P}_1\|\boldsymbol{Q}_1\right)\right)^n$$

Thus,

$$2-2\exp\left(-\frac{nH^2\left(\boldsymbol{P}_1\|\boldsymbol{Q}_1\right)}{2}\right)\leq\|\boldsymbol{P}^{1:n}-\boldsymbol{Q}^{1:n}\|_{\mathrm{H}}^2\leq nH^2\left(\boldsymbol{P}_1\|\boldsymbol{Q}_1\right)$$

We often use the Hellinger affinity as the notation is more convenient at times, which is defined as

$$\text{Aff}(\boldsymbol{P}, \boldsymbol{Q}) = \int_{\mathcal{X}} \sqrt{p(x)q(x)} d\nu(x).$$

Notice that $\frac{1}{2}\|\boldsymbol{P} - \boldsymbol{Q}\|_{\text{H}}^2 = 1 - \text{Aff}(\boldsymbol{P}, \boldsymbol{Q})$. Putting everything together, we have the following lemma:

Lemma 1.1. *If Aff*$(\boldsymbol{P}^{1:n}, \boldsymbol{Q}^{1:n}) = 1 + o(1/n)$*, then* $\|\boldsymbol{P}^{1:n} - \boldsymbol{Q}^{1:n}\|_{TV}^2 = o(1)$ *as* $n \to \infty$*; and if Aff*$(\boldsymbol{P}^{1:n}, \boldsymbol{Q}^{1:n}) = 1 - \omega(1/n)$*, then* $\|\boldsymbol{P}^{1:n} - \boldsymbol{Q}^{1:n}\|_{TV}^2 \to 1$ *as* $n \to \infty$.

1.2.1 Proof for Lemma ??

Proof. Let $\mu(w) \propto \min(P_w, Q_w)$ for $w \in \mathcal{W}$ and $\nu(w) \propto \max(0, P_w - Q_w)$ for $w \in \mathcal{W}$. With $\mathcal{A} := \{w \in \mathcal{W} : P_w \geq Q_w\}$ and $p = \sum_{w \in \mathcal{W}} \min(P_w, Q_w)$, we have $\mu(w) = \min(P_w, Q_w)/p$ for $w \in \mathcal{W}$, and $\nu(w) = (P_w - Q_w)/(1-p)$ for $w \in \mathcal{A}$ and 0 otherwise. Then, for any $w \in \mathcal{A}, p \cdot \mu(w) + (1-p) \cdot \nu(w) = Q_w + P_w - Q_w = P_w$ and for any $w \notin \mathcal{A}, p \cdot \mu(w) + (1-p) \cdot \nu(w) = P_w + 0 = P_w$. Therefore, $w \sim \boldsymbol{P}$. □

1.3 Proof for the Detection Scheme

First, we assume no modification was made to the generated text; we will discuss the substitution attack in subsequent sections. Recall that as we restrict our attention to the m green tokens, we can reformulate the detection scheme into a hypothesis-testing problem:

H_0: the text $\tilde{w}_{1:n}$ is independent of the decoder; i.e. $\zeta_t \overset{\text{i.i.d.}}{\sim} U[0,1]$ for $t = 1, ..., m$.

H_1: the text $\tilde{w}_{1:n}$ is generated from the decoder; $\zeta_t | (\boldsymbol{P}_t, \mathcal{G}_t) \sim U[0, P_{t,\mathcal{G}_t}]$ for $t = 1, ..., m$.

1.3.0.1 Order Statistics. We first discuss the simplest case, where each $P_{t,\mathcal{G}_t}$ is equal and bounded away from 1. Then we progressively generalize the proof. In this section, let

$\zeta_{1:m} = (\zeta_1, ..., \zeta_m)$ and $\zeta_{(m)} = \max(\zeta_1, ..., \zeta_m)$.

Formally, we assume that there exists some constant $\delta > 0$ such that $P_{t,\mathcal{G}_t} = 1 - \delta$ for each $t = 1, ..., m$. Under this assumption, Neyman-Pearson Lemma implies that the likelihood ratio test is the uniformly most powerful (UMP). In particular, the likelihood ratio test is given by

$$\Lambda(\zeta_{1:n}) = \frac{\prod_{t=1}^{m} \mathbb{1}_{\zeta_j \in [0, P_{t,\mathcal{G}_t}]} / P_{t,\mathcal{G}_t}}{\prod_{t=1}^{m} \mathbb{1}_{\zeta_t \in [0,1]}} = (1-\delta)^{-m} \mathbb{1}_{\zeta_{(m)} \leq 1-\delta}$$

where $\mathbb{1}$ denotes the indicator function. This likelihood-ratio test provides the decision rule as follows: for some critical value c,

- if $\Lambda(\zeta_{1:n}) \geq c$, then reject H_0;
- if $\Lambda(\zeta_{1:n}) < c$, then accept H_0.

This is equivalent to the following decision rule: for some critical value c',

- if $\zeta_{(m)} \leq c'$, then reject H_0;
- if $\zeta_{(m)} > c'$, then accept H_0.

Thus, to construct the UMP test, we need to find a critical value c' such that:

$$\mathbb{P}_{H_0}(\zeta_{(m)} \leq c') = \alpha$$

where α is the significance level. Since $\zeta_{(m)}$ is the maximum of m i.i.d. uniform random variables, we have

$$\mathbb{P}_{H_0}(\zeta_{(m)} \leq c') = \prod_{t=1}^{m} [\mathbb{P}_{H_0}(\zeta_t \leq c')]^m = (c')^m.$$

Hence, we can choose $c' = \alpha^{1/m}$ to obtain the desired significance level. The power of the

test is given by

$$\mathbb{P}_{H_1}(\zeta_{(m)} \leq c') = \prod_{t=1}^{m} [\mathbb{P}_{H_1}(\zeta_t \leq c')]^m = \begin{cases} \left(\frac{c'}{1-\delta}\right)^m = \frac{\alpha}{(1-\delta)^m} & \text{for } c' \leq 1-\delta \\ 1 & \text{for } 1-\delta < c' \leq 1. \end{cases}$$

For fixed α and δ, as m increases, $c' = \alpha^{1/m}$ will be greater than $1-\delta$, which leads the power to be equal to 1. In the case where $P_{t,\mathcal{G}_t} \leq 1-\delta$ for all t, the above argument gives a lower bound on the power of the same test.

1.3.0.2 Text Modifications. Before we discuss more general settings for $P_{t,\mathcal{G}_t}$, we first discuss the case where the user can modify the generated text only via substituting some tokens. In particular, we assume that the user can substitute any token in the generated text with any other token without the knowledge of the green lists or ζ_t's. When a token is substituted, the corresponding uniform random variable ζ becomes independent of whether the new token falls into the corresponding green list. Under this circumstance, the statistical signal of the watermark becomes sparse, leading us to consider an alternative hypothesis where a fraction ε_m of the ζ_t's still present signals of the watermark. Formally, we have the following hypothesis:

$H_1^{(\text{mix})}$: the text $\tilde{w}_{1:n}$ is first generated from the decoder, and then modified by substitution attacks described above; that is, $\zeta_t|(\boldsymbol{P}_t, \mathcal{G}_t) \sim (1-\varepsilon_m)U[0,1] + \varepsilon_m U[0, P_{t,\mathcal{G}_t}]$ for $t = 1, ..., m$.

Throughout our discussion, we treat $m\varepsilon_m$ as an integer for simplicity.

Suboptimality of the Order Statistics. We apply the same test as above, where we reject the null hypothesis if and only if $\zeta_{(m)} \leq c'$. Mimicking the above analysis, we can

reach the following lemma:

Lemma 1.2. *In the presence of the substitution attack described above, the power of this test is given by*

$$\mathbb{P}_{H_1}(\zeta_{(m)} \leq c') = \begin{cases} \frac{\alpha}{(1-\delta)^{m\cdot\varepsilon_m}} & \text{for } c' \leq 1-\delta \\ \alpha^{1-\varepsilon_m} & \text{for } 1-\delta < c' \leq 1. \end{cases}$$

Proof. The proof is similar to the above analysis. When $c' \leq 1-\delta$,

$$\mathbb{P}_{H_1}(\zeta_{(m)} \leq c') = \prod_{t=1}^{m} [\mathbb{P}_{H_1}(\zeta_t \leq c')]^m = (c')^{m(1-\varepsilon_m)} \cdot \left(\frac{c'}{1-\delta}\right)^{m\varepsilon_m} = \frac{\alpha}{(1-\delta)^{m\varepsilon_m}}$$

where the last equality follows from the fact that $c' = \alpha^{1/m}$. On the other hand, when $c' > 1-\delta$, we can assume without loss of generality that only the first $m\varepsilon_m$ of the ζ_t's presents the watermark signal. As these ζ's are bounded above by $1-\delta$, we have

$$\begin{aligned} \mathbb{P}_{H_1}(\zeta_{(m)} \leq c') &= \prod_{t=1}^{m\varepsilon_m} [\mathbb{P}_{H_1}(\zeta_t \leq c')]^m \cdot \prod_{t=m\varepsilon_m+1}^{m} [\mathbb{P}_{H_1}(\zeta_t \leq c')]^m \\ &= (c')^{m(1-\varepsilon_m)} = \alpha^{1-\varepsilon_m}. \end{aligned}$$

□

Notice that the power of the test is now upper bounded by $\alpha^{1-\varepsilon_m}$. This means, unless $\varepsilon_m \to 1$, i.e., the fraction of modification becomes negligible as the length of the text increases, the power of the test will be suboptimal, leading to the next proposition.

Proposition 1.3. *For a given significance level $\alpha < 1$, if ε_m does not converge to 1, then the power of the test based on the order statistic $\zeta_{(m)}$ is bounded away from 1.*

Undetectability. We now lift the assumption that all $P_{t,\mathcal{G}_t}$'s are equal in our subsequent discussion.

Proof of Theorem **??**. As observed earlier, we can assume without loss of generality that $P_{t,\mathcal{G}_t} = P_{\mathcal{G}} = m^{-r}$ for all t. By Lemma 1.1, it suffices to show that the Hellinger affinity between $U[0,1]$ and $(1-\varepsilon_m)U[0,1] + \varepsilon_m U[0, P_{\mathcal{G}}]$ behaves asymptotically as $1 + o(1/m)$. Let $f(y)$ and $g(y)$ be the densities of $U[0,1]$ and $U[0, P_{\mathcal{G}}]$ respectively. As $f(y) = 1$ and $g(y) = P_{\mathcal{G}}^{-1}\mathbb{1}_{x\in[0,P_{\mathcal{G}}]}$ on $[0,1]$, the Hellinger affinity is given by

$$\begin{aligned}\mathbb{E}_0\sqrt{f(y)((1-\varepsilon_m)f(y)+\varepsilon_m g(y))} &= \mathbb{E}_0\sqrt{(1-\varepsilon_m)+\varepsilon_m g(y)} \\ &= \mathbb{E}_0\sqrt{1+\varepsilon_m(g(y)-1)}\end{aligned}$$

where $\mathbb{E}_0$ denotes the expectation under the null hypothesis. As $0 < r, p < 1$ and $2p - r > 1$, we must have $r < p$. Hence, $\varepsilon_m g(y) \le \varepsilon_m p(G)^{-1} = m^{-p}m^r \to 0$. This implies that for large enough m, $x := \varepsilon_m(g(y)-1)$ satisfies

$$1 + \frac{x}{2} - \frac{x^2}{8} \le \sqrt{1+x} \le 1 + \frac{x}{2}.$$

Hence, as $\mathbb{E}_0 g(y) = 1$, we have

$$1 - \mathbb{E}_0\frac{(\varepsilon_m(g(y)-1))^2}{8} \le \sqrt{1+\varepsilon_m(g(y)-1)} \le 1.$$

It remains to show that $\mathbb{E}_0(\varepsilon_m(g(y)-1))^2 = o(1/m)$: again, as $\mathbb{E}_0 g(y) = 1$,

$$\begin{aligned}
\mathbb{E}_0(\varepsilon_m(g(y)-1))^2 &= \varepsilon_m^2 \mathbb{E}_0((g(y)-1))^2 \\
&= \varepsilon_m^2 \left(\mathbb{E}_0(g(y)^2 - 1)\right) \\
&= \varepsilon_m^2 \left(\mathbb{E}_0 p(G)^{-2} \mathbb{1}_{x\in[0,P_{\mathcal{G}}]} - 1\right) \\
&= \varepsilon_m^2 \left(P_{\mathcal{G}}^{-1} - 1\right) \\
&= m^{-2p}(m^r - 1) = o(1/m)
\end{aligned}$$

where the last equality follows from the assumption that $2p - r > 1$. □

Detection Boundary.

Proof of Theorem **??**. We first prove the first part of the theorem. In this case, we can restrict our attention to the case where each $P_{t,\mathcal{G}_t} = P_{\mathcal{G}} = 1 - m^{-q}$ without loss of generality, as smaller $P_{t,\mathcal{G}_t}$ only present more watermark signals. By Lemma 1.1, it suffices to show that the Hellinger affinity between $U[0,1]$ and $(1-\varepsilon_m)U[0,1] + \varepsilon_m U[0, P_{\mathcal{G}}]$ behaves asymptotically as $1 - \omega(1/m)$. Let $f(y)$ and $g(y)$ be the densities of $U[0,1]$ and $U[0, P_{\mathcal{G}}]$ respectively. As $\varepsilon_m(g(y)-1) = m^{-p} \cdot m^{-q}/(1-m^{-q}) \to 0$, similar to the last proof, it is enough to investigate the behavior of $\mathbb{E}_0(\varepsilon_m(g(y)-1))^2$:

$$\begin{aligned}
\mathbb{E}_0(\varepsilon_m(g(y)-1))^2 &= \varepsilon_m^2 \left(\mathbb{E}_0(g(y)^2 - 1)\right) \\
&= \varepsilon_m^2 \left(P_{\mathcal{G}}^{-1} - 1\right) \\
&= m^{-2p} \cdot \frac{m^{-q}}{1 - m^{-q}} = \omega(1/m)
\end{aligned} \tag{2}$$

because $q + 2p < 1$. Hence, H_0 and $H_1^{(\text{mix})}$ separate asymptotically. To show that the alternative hypothesis can be reliably detected using the LRT, it is sufficient to show that

the sum of type I and type II error probabilities tends to 0 as $m \to \infty$. Since the proofs are similar, we present the argument for the type I error under the null hypothesis. Let $\ell = \log(1 + \varepsilon_m(g(y) - 1))$ and $L_m = m\ell$. It suffices to show

$$\mathbb{E}_0 L_m \to -\infty \quad \text{and} \quad \frac{\mathrm{Var}_0(L_m)}{[\mathbb{E}_0 L_m]^2} \to 0$$

Let $x := \varepsilon_m(g(y) - 1)$ and use the fact that $\log(1 + x) \leq x - x^2/4$ for $x \in (-1, 1]$. For large enough m, we have

$$\mathbb{E}_0 \ell \leq \varepsilon_m \mathbb{E}_0(g(y) - 1) - \frac{\varepsilon_m^2}{4}\mathbb{E}_0(g(y) - 1)^2 = -\frac{\varepsilon_m^2}{4}\mathbb{E}_0(g(y) - 1)^2 = -\omega(1/m)$$

where the last equality follows from Equation (2). This gives us $\mathbb{E}_0 L_m \to -\infty$. On the other hand, using the fact that $\log^2(1 + x) \leq 2x^2$ for $x \in (-0.5, 1]$, we have for large enough m,

$$\mathrm{Var}_0 \, \ell \leq \mathbb{E}_0 \ell^2 \leq 2\varepsilon_m^2 \mathbb{E}_0(g(y) - 1)^2 \leq 8|\mathbb{E}_0 \ell|$$

This gives us $\mathrm{Var}_0 \, L_m / [\mathbb{E}_0 L_m]^2 \to 0$.

We now prove the second part of the theorem. In this case, we can restrict our attention to the case where each $P_{t,\mathcal{G}_t} = P_{\mathcal{G}} = 1 - m^{-q}$. Following the same steps, we have

$$\mathbb{E}_0(\varepsilon_m(g(y) - 1))^2 = \varepsilon_m^2(P_{\mathcal{G}}^{-1} - 1) = m^{-2p}\frac{m^{-q}}{1 - m^{-q}} = o(1/m)$$

as $q + 2p > 1$. Hence, H_0 and $H_1^{(\mathrm{mix})}$ merge asymptotically. □

Sum Test.

Proof of Proposition **??**. For both parts of the proposition, we can assume without loss of generality that each $P_{t,\mathcal{G}_t} = P_{\mathcal{G}} = 1 - m^{-q}$.

As for the first part, the proofs for showing that both type I and type II error probabilities tend to 0 are very similar, so we present the argument for the type II error under the alternative hypothesis as it involves slightly more technicality. Let $s = \sum_{i=1}^{m} \zeta_i$ and $c' = \mathbb{E}_0 s - m^{1-(p+q)}$. Then

$$\mathbb{E}_1 s = m\mathbb{E}_1\zeta_1 = m\left(\frac{1-\varepsilon_m}{2} + \frac{\varepsilon_m \cdot P_{\mathcal{G}}}{2}\right) = \frac{m}{2} - \frac{m^{1-(p+q)}}{2}$$

and

$$\operatorname{Var}_1 s = m \operatorname{Var}_1 \zeta_1 = O(m)$$

By Chebyshev's inequality, we have

$$\mathbb{P}_1(s \le c') \le \mathbb{P}_1(|s - \mathbb{E}_1 s| \le c' - \mathbb{E}_1 s) = \mathbb{P}_1(|s - \mathbb{E}_1 s| \le m^{1-(p+q)}/4) \le \frac{4 \operatorname{Var}_1 s}{m^{2-2(p+q)}} = o(1)$$

where the last equality follows from the fact that $q + p < 1/2$ and $\operatorname{Var}_1 s = O(m)$.

Now, for the second part of the proposition, it suffices to observe that $\mathbb{P}_1(s \le \mathbb{E}_0 s)$ is bounded away from 1. As $p+q > 1/2, \mathbb{E}_0 s - \mathbb{E}_1 s = m^{1-(p+q)}/2 <= m^{1/2}$ and $\operatorname{Var}_1 s = O(m)$, which implies that as $m \to \infty, \mathbb{E}_0 s$ is within one standard deviation of the center of the distribution of s under the alternative hypothesis. Hence, the power of the test must be bounded away from 1. □

Higher Criticism.

Proof of Theorem **??**. Under the null hypothesis, HC_m^* is equal to the extreme value of a

normalized uniform empirical process in distribution, which satisfies

$$\frac{HC_m^*}{\sqrt{2\log\log m}} \to 1, \quad \text{in probability},$$

so $\mathbb{P}_0(HC_m^* \geq \sqrt{(2+\delta)\log\log m}) \to 0$ as $m \to \infty$; see Theorem 1.1. in Donoho and Jin [2004]. Before we show that $\mathbb{P}_1(HC_m^* \leq \sqrt{(2+\delta)\log\log m}) \to 0$, we present an equivalent form of the higher criticism statistic. Consider the empirical cumulative distribution function

$$F_m(x) = \frac{1}{m}\sum_{i=1}^{m} \mathbb{1}_{\zeta_i \leq x},$$

and its standardized version

$$W_m(x) = \sqrt{m} \cdot \frac{F_m(x) - x}{\sqrt{x(1-x)}}.$$

Then, we observe that the higher criticism statistic satisfies

$$HC_m^* = \max_{1\leq t\leq m} \sqrt{m} \cdot \left(\frac{t/m - \zeta_{(t)}}{\zeta_{(t)}(1-\zeta_{(t)})}\right) = \max_{1\leq t\leq m} \sqrt{m} \cdot \left(\frac{F_m(\zeta_{(t)}) - \zeta_{(t)}}{\zeta_{(t)}(1-\zeta_{(t)})}\right) = \sup_{x\in[0,1]} W_m(x).$$

If ζ_t's are i.i.d. uniform random variables, then $\mathbb{E}F_m(x) = x$ for all $x \in [0,1]$; while if ζ_t's are independent uniform random variables in $[0, P_{t,\mathcal{G}_t}]$, where $P_{t,\mathcal{G}_t} \leq 1 - m^{-q}$, then

$\mathbb{E}F_m(1-m^{-q}) = 1$. Hence, setting $w = 1 - m^{-q}$,

$$\begin{aligned}
\mathbb{E}_1 W_m(x) &= \sqrt{m} \cdot \left(\frac{\mathbb{E}_1 F_m(x) - x}{\sqrt{x(1-x)}} \right) \\
&= \sqrt{m} \cdot \left(\frac{\frac{1}{m}\sum_{i=1}^m \mathbb{E}_1 \mathbb{1}_{\zeta_i \le x} - x}{\sqrt{x(1-x)}} \right) \\
&= \sqrt{m} \cdot \left(\frac{\varepsilon_m(1-x)}{\sqrt{x(1-x)}} \right) \\
&= \sqrt{m} \cdot \left(\frac{m^{-p} m^{-q}}{\sqrt{(1-m^{-q})m^{-q}}} \right) \\
&= \sqrt{\left(\frac{m^{1-2p-q}}{1-m^{-q}} \right)} = \Omega(n^{\gamma})
\end{aligned}$$

for some $0 < \gamma \le (1-(2p+q))/2$ and

$$\frac{\mathbb{E}_1 W_m(x)}{\sqrt{(2+\delta)\log\log m}} \to \infty.$$

Meanwhile,

$$\mathrm{Var}_1\, W_m(x) = \frac{F_m(x)(1-F_m(x))}{x(1-x)} = O(1).$$

Hence, as $HC_m^* \ge W_m(x)$ for all $x \in [0,1]$, by Chebyshev's inequality, for large enough constant C, C',

$$\begin{aligned}
\mathbb{P}_1(HC_m^* \le \sqrt{(2+\delta)\log\log m}) &\le \mathbb{P}_1(W_m(x) \le \sqrt{(2+\delta)\log\log m}) \\
&\le C\frac{\mathrm{Var}_1\, W_m(x)}{[\mathbb{E}_1 W_m(x)]^2} \le C' n^{-2\gamma} \to 0
\end{aligned}$$

which completes the proof. □

1.4 Extension to the Soft Green/Red List Watermark

When the generated text is expected to be very long, we can adapt our proposed scheme to the soft watermark; this may further mitigate the possible distortion during the text generation process when pseudorandomness based on the previous tokens is used. To do this, we simply replace Q_w with the method described in Equation (1). In this case, we will not zero out the probability of sampling a token from the red list $\mathcal{R}$ in Q_w. Thus, even in the unfortunate situation where the last k tokens result in a very low pseudorandom number, there is still a slight possibility of choosing a word from the "red list." However, this flexibility can diminish the method's ability to detect the watermark: firstly, we can rewrite Lemma **??** as follows:

Lemma 1.4. *Let $\boldsymbol{P}_t$ be the original token distribution at step t and w be the next token sampled from the above decoding scheme. Then the conditional distribution of ζ_t, given $w \in \mathcal{G}_t$, is uniformly distributed over*

$$\left[0, \frac{1 + (\mathrm{e}^{\delta} - 1)P_{t,G_t}}{\mathrm{e}^{\delta}} = P_{t,G_t} + \frac{1 - P_{t,G_t}}{\mathrm{e}^{\delta}}\right].$$

This revised statement suggests that when a green token is chosen, the watermark signal can be much weaker than in the original scheme. Secondly, the symmetry observed in Remark **??** should be revised accordingly. If the next token w is red, then the conditional distribution of ζ_t can be either a standard uniform random variable (if the first sample w is not rejected) or a uniform random variable in the interval $\left[P_{t,\mathcal{G}_t} + \frac{1-P_{t,\mathcal{G}_t}}{\mathrm{e}^{\delta}}, 1\right]$ (if the first sample w is rejected). Hence, the proportion of signals coming from the red tokens will be much smaller than in the original scheme. Nonetheless, it is evident that analogous asymptotic guarantees to Proposition **??** and Theorem **??** can be established with appropriate assumptions.

1.5 Speculative Decoding

Speculative decoding [Leviathan et al., 2023, Chen et al., 2023] is an algorithm originally designed to speed up sampling text from a large target language model by using a smaller, faster model, and has been widely adopted in production. Essentially, for a lookahead parameter L, the small draft model generates L tokens ahead at each step, which the larger target model then evaluates in parallel, either accepting the proposed sequence up to the first rejected token or falling back to standard autoregressive generation if the first token is rejected. This process is formalized in Algorithm 1.

Algorithm 1 Speculative Sampling with Draft Model

Input: Target model $\mathcal{M}_P$, draft model $\mathcal{M}_Q$, random variables $\zeta_1, ..., \zeta_L \in [0, 1]$, prompt x
Output: Accepted tokens sequence y
for $t = 1$ to L **do**
  Compute $\boldsymbol{Q}_t = \mathcal{M}_P(\cdot|x, w_1, ..., w_t - 1)$ and sample w_t from $\boldsymbol{Q}_t$
end for
Compute $\boldsymbol{P}_t = \mathcal{M}_P(\cdot|x, w_1, ..., w_t)$ for $t = 1, ..., L$ in parallel
for $t = 1$ to L **do**
  if $\zeta_t \cdot Q_{t,w} > P_{t,w}$ **then**
    Sample w'_t from the normalized excess distribution $\propto \max(0, \boldsymbol{P}_t - \boldsymbol{Q}_t)$
    return $(w_{1:t-1}, w'_t)$
  end if
end for
return $(w_{1:L})$

1.5.0.1 Connections to Maximal Coupling

When the lookahead parameter L equals 1, the algorithm simplifies to a basic rejection sampling scheme, as shown in Algorithm 2. Now we prove that this process is equivalent to the maximal coupling defined in Algorithm **??** if ζ is a standard uniform random variable. Let `Accept` be the event where we keep the initially sampled token. If we resample w from the normalized excess distribution $\propto \max(0, \boldsymbol{P}_t - \boldsymbol{Q}_t)$, the process is equivalent to that of the maximal coupling in Algorithm **??**. Hence, we first see that $w|$`Accept` follows the normalized overlap distribution $\sim \min(\boldsymbol{P}, \boldsymbol{Q})$:

with a slight abuse of notation, we use $\mathbb{P}[w]$ to denote the probability of sampling the token $w \in \mathcal{W}$ by following Algorithm **??**; then

$$\mathbb{P}[w|\texttt{Accept}] = Q_w \cdot \mathbb{P}[\zeta \leq \min(1, P_w/Q_w)] = Q_w \cdot \min(1, P_w/Q_w) = \min(P_w, Q_w)$$

Now, it remains to show that $\mathbb{P}[\texttt{Accept}] = \sum_{w \in \mathcal{W}} \min(P_w, Q_w)$:

$$\mathbb{P}[\texttt{Accept}] = \mathbb{P}_{w \sim \boldsymbol{Q}}[\zeta \cdot Q_w > P_w] = \sum_{w \in \mathcal{W}} Q_w \cdot \min(\frac{P_w}{Q_w}, 1) = \sum_{w \in \mathcal{W}} \min(P_w, Q_w) = p$$

Algorithm 2 One Step Speculative Sampling for Next Token Generation

Input: Token distributions $\boldsymbol{P}, \boldsymbol{Q}$, random variable $\zeta \in [0, 1]$
Output: Sampled token w
Independently sample $w \sim \boldsymbol{Q}$
if $\zeta \cdot Q_w > P_w$ **then**
  Sample w' from the normalized excess distribution $\propto \max(0, \boldsymbol{P} - \boldsymbol{Q})$
  $w \leftarrow w'$
end if
return w

1.5.0.2 Speculative Decoding as Post-Processing

To investigate the performance of the watermarking schemes under targeted modifications, we consider the following speculative decoding setup. We view the NTP generated by a watermarked decoder as a probability vector $\boldsymbol{Q}$ conditioned on the previous context. This context influences both the original language model's NTP generation and the pseudorandom variable $\tilde{\zeta}$ that determines the watermark signal. Concretely, let $\tilde{\boldsymbol{P}}$ denote the NTP generated by the original draft language model. For the Gumbel-max watermarking scheme, $\boldsymbol{Q}$ is given by

$$Q_w = \mathbb{1}_{\{w = \text{argmax}_{w' \in \mathcal{W}} \log U_{w'}/\tilde{P}'_w\}}$$

where $U_{w'}$ is the pseudorandom variables determined by the prevoius tokens. For our proposed watermarking scheme, $\boldsymbol{Q}$ is given by

$$Q_w = \mathbb{1}_{\{\tilde{\zeta} \leq \tilde{P}_{\mathcal{G}_t}\}} \frac{\tilde{P}_w}{\tilde{P}_{\mathcal{G}_t}} + \mathbb{1}_{\{\tilde{\zeta} > \tilde{P}_{\mathcal{G}_t}\}} \frac{\tilde{P}_w}{1 - \tilde{P}_{\mathcal{G}_t}}$$

where $\tilde{P}_{\mathcal{G}_t} = \sum_{w' \in \mathcal{G}_t} \tilde{P}_{w'}$, $\mathcal{G}_t$ is the green list determined by the previous tokens and $\tilde{\zeta}$ is the pseudorandom variable used for watermarking.

We assume the target model produces an NTP $\boldsymbol{P}$ that, given the previous tokens, is independent of the pseudorandom variables used for watermarking the draft model. This means the target model makes judgments based solely on $\boldsymbol{P}$, without knowledge of the watermark signal. Simply applying Algorithm 2 to $\boldsymbol{P}, \boldsymbol{Q}$ and an independent standard uniform variable ζ would erase the watermark signal, as the marginal distribution of the sampled token w would match the original NTP $\boldsymbol{P}$. Therefore, we model the target model as a "lazy" editor by modifying the rejection sampling condition to $0.5 \cdot \zeta \cdot Q_w > P_w$. This means the target model only accepts the draft model's suggestion when it has reasonable confidence in generating the same token. For instance, if $Q_w = 1$, the target model accepts the suggestion only if it has at least 50% confidence in generating that token. This assumption simulates a human editor who prefers making minimal modifications to the draft model's suggestions.

1.6 More Experimental Results

1.6.1 Additional Simulation Studies

Complementary to Figure **??**, Figure 1 also shows that the sum test tends to outperform the higher criticism when m is relatively small.

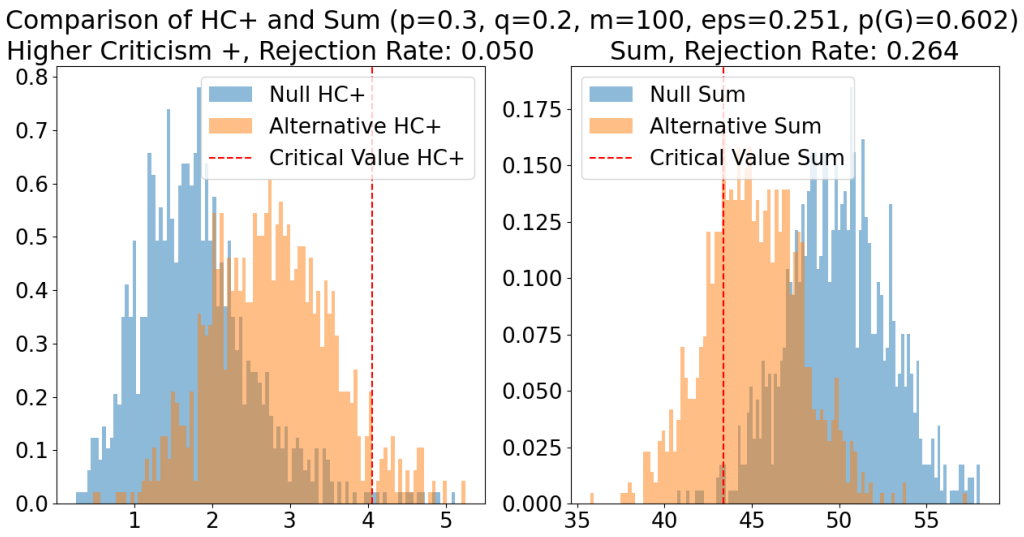

Figure 1: The histograms of the test statistics under the null and alternative hypotheses for the sum test and the higher criticism when $m = 100$. The test statistics are computed from the simulation under both the null and alternative hypotheses 2000 times. The critical value is determined by the quantile of the simulated histogram of the test statistic under the null hypothesis. The rejection rate, or the power, of each test, is computed by computing, under the alternative hypothesis, the proportion of the test statistics that are on the other side of the critical value. See also Figure **??**.

1.6.2 Further Implementation Details on Language Model Experiments

In our experiments, we prompt two family of instruction fine-tuned models with their respective default instruction prompting templates: Meta's `LLaMA` model family [Dubey et al., 2024] and Microsoft's `Phi-3` family [Abdin et al., 2024]. For each experiment, we embed watermarks to `Llama-3.2-1B-Instruct` (1B parameters) and `Phi-3-mini-4k-instruct` (3.8B parameters). For both models, we apply Flash Attention to optimize for faster inference [Saha and Ye, 2024]. For the two datasets, we use the same 200 samples from both datasets as used in [Tu et al., 2023]. These sub-samples of the two datasets provide short input instructions but should yield relatively long answers: for ELI5, the average length of the input question is 41.04 words while the average length of the reference answer is 236.6 words; similarly, for FinQA, the average length of the input question is 13.67 while the average length of the reference answer is 251.13 words. Our prompts follow the chat

template of each respective model.

For the speculative decoding setup, we choose `Llama-3.2-3B-Instruct` (3B parameters) as the target model for the `LLaMA` family, and `Phi-3-medium-4k-instruct` (14 B parameters) for the `Phi-3` family.

For our sum test statistic, due to precision errors and computational complexity in calculating the exact Irwin-Hall distribution, when computing the p-values, we use normal approximation when $n \geq 15$. Unless otherwise specified, we flag a text as watermarked if the p-value is lower than $\alpha = 0.01$, ensuring a false positive rate to be less than 0.01. For higher criticism, instead of computing the p-values, we precompute the quantile of the simulated histogram of the test statistic under the null hypothesis to determine the empirical critical value at the significance level $\alpha = 0.01$.

1.6.3 Ablation Study on the Parameters

Table 1 present the analogous results to Table **??** for the `Llama-3.2-1B-Instruct` model. The results are consistent with the main results, showing that our proposed watermarking scheme outperforms the Gumbel-max method in terms of the text distortion, and more competitive TPR than the green/red list and DiPMark methods.

Table 2 and Table 3 present the analogous results with a wider range of parameters. Table 3 contains the results for the Llama model, while Table 2 contains the results for the Phi model. From both tables, we observe a positive correlation between the distortion of the generation (lower S-BERT scores) and the detection power (higher TPR) for the green/red list method, echoing existing theory and empirical findings [Kirchenbauer et al., 2023a, Fernandez et al., 2023, Piet et al., 2023].

Table 4 presents the analogous results to Table **??** for the Llama model. Here, we observe

Table 1: Comparison of the S-BERT similarity scores and the true positive rates (TPR) among four watermarking schemes: (1) the Gumbel-max watermark [Aaronson and Kirchner, 2023], (2) the green/red list watermark [Kirchenbauer et al., 2023a], (3) the DiPmark method Wu et al. [2023], and (4) our proposed one.

Model	Data	k	Metric		Gumbel-max	Green/red list	DiPmark	Ours
Llama-3.2-1B-Instruct	FinQA	2	S-BERT	mean	0.7945	0.8038	0.8098	0.8136
				median	0.8175	0.8292	0.8333	0.8347
			TPR		1.0000	0.7550	0.9200	**0.9500**
			TPR aug.		0.9950	0.5150	0.7350	**0.8550**
			TPR para.		0.9800	0.4100	0.5400	**0.6800**
		4	S-BERT	mean	0.8020	0.8270	0.8086	0.8117
				median	0.8344	0.8458	0.8328	0.8426
			TPR		0.9850	0.8300	0.9700	**0.9750**
			TPR aug.		0.9850	0.4500	0.7550	**0.8400**
			TPR para.		0.9500	0.2650	0.4300	**0.6450**
	ELI5	2	S-BERT	mean	0.7171	0.7195	0.7250	0.7351
				median	0.7342	0.7315	0.7393	0.7429
			TPR		1.0000	0.9100	0.9950	**1.0000**
			TPR aug.		1.0000	0.6650	0.9100	**1.0000**
			TPR para.		1.0000	0.4500	0.8600	**0.9200**
		4	S-BERT	mean	0.7302	0.7226	0.7250	0.7315
				median	0.7447	0.7433	0.7346	0.7433
			TPR		1.0000	0.9700	**1.0000**	**1.0000**
			TPR aug.		1.0000	0.6450	0.9350	**0.9900**
			TPR para.		1.0000	0.4000	0.7800	**0.9200**

that for the Llama model, the Gumbel-max method has a higher rate of repeated tokens compared to all other methods.

Table 5 presents the analogous results to Table **??** for the Llama model. The results show that our proposed watermarking scheme has a lower average rejection rate compared to the Gumbel-max method, and this is consistent with the main results.

Table 6 presents the analogous results to Table **??** for the Llama model. The results show higher true positive rate for the sum test compared to the higher criticism when the number of tokens used to detect the watermark is relatively small, and this consistent with the main results.

Table 2: Comparison of the S-BERT similarity scores and the true positive rates (TPR) of the two watermarking schemes: green/red list Kirchenbauer et al. [2023a] and our proposed scheme for the Phi-3-mini-4k-instruct (3.8B).

Data	k	Metric		Green/red list $\delta = 2$ $\gamma = 0.25$	Green/red list $\delta = 2$ $\gamma = 0.5$	Green/red list $\delta = 1$ $\gamma = 0.25$	Green/red list $\delta = 1$ $\gamma = 0.5$	Our method $\gamma = 0.5$
FinQA	2	S-BERT	mean	0.7887	0.8001	0.8269	0.8167	0.7963
			median	0.8229	0.8404	0.8598	0.8472	0.8382
		TPR		0.9700	0.9700	0.5900	0.6550	0.9750
		TPR aug.		0.9150	0.8850	0.3250	0.3500	0.8700
	4	S-BERT	mean	0.7957	0.7967	0.8305	0.8233	0.8023
			median	0.8241	0.8277	0.8528	0.8494	0.8288
		TPR		0.9650	0.9600	0.7850	0.7750	0.9500
		TPR aug.		0.8750	0.8900	0.4450	0.5150	0.8300
ELI5	2	S-BERT	mean	0.7022	0.6962	0.7084	0.7178	0.7161
			median	0.7175	0.7060	0.7264	0.7239	0.7290
		TPR		1.0000	1.0000	0.7800	0.9100	0.9900
		TPR aug.		0.9900	0.9850	0.4900	0.6500	0.9350
	4	S-BERT	mean	0.7030	0.7161	0.7060	0.7084	0.7185
			median	0.7076	0.7299	0.7214	0.7236	0.7344
		TPR		1.0000	0.9950	0.8850	0.8850	1.0000
		TPR aug.		0.9500	0.9850	0.5450	0.5750	0.9550

Table 3: Same as Table 2, but for the Llama model.

Data	k	Metric		Green/red list $\delta = 2$ $\gamma = 0.25$	Green/red list $\delta = 2$ $\gamma = 0.5$	Green/red list $\delta = 1$ $\gamma = 0.25$	Green/red list $\delta = 1$ $\gamma = 0.5$	Our method $\gamma = 0.5$
FinQA	2	S-BERT	mean	0.7962	0.8025	0.8199	0.8038	0.8136
			median	0.8144	0.8290	0.8391	0.8292	0.8347
		TPR		0.9900	0.9950	0.6750	0.7550	0.9500
		TPR aug.		0.9350	0.9550	0.4450	0.5150	0.8550
	4	S-BERT	mean	0.8059	0.7940	0.8247	0.8270	0.8117
			median	0.8318	0.8127	0.8381	0.8458	0.8426
		TPR		0.9750	0.9850	0.8300	0.8300	0.9750
		TPR aug.		0.9250	0.9300	0.5150	0.4500	0.8400
ELI5	2	S-BERT	mean	0.7077	0.7040	0.7196	0.7195	0.7351
			median	0.7181	0.7267	0.7382	0.7315	0.7429
		TPR		1.0000	1.0000	0.9250	0.9100	1.0000
		TPR aug.		0.9950	0.9900	0.7150	0.6650	0.9650
	4	S-BERT	mean	0.7169	0.7008	0.7279	0.7226	0.7315
			median	0.7385	0.7243	0.7514	0.7433	0.7433
		TPR		1.0000	1.0000	0.9400	0.9700	1.0000
		TPR aug.		1.0000	0.9950	0.6450	0.6450	0.9900

Table 4: Same as Table **??**, but for the Llama model.

Model	Dataset	k	Method	Repeated (%)
Llama-3.2-1B-Instruct	FinQA	2	Gumbel-max	25.16%
			Green/red list	19.39%
			DiPMark	21.03%
			Ours	21.29%
		4	Gumbel-max	7.81%
			Green/red list	5.94%
			DiPMark	5.85%
			Ours	6.17%
	ELI5	2	Gumbel-max	25.25%
			Green/red list	18.81%
			DiPMark	19.58%
			Ours	19.50%
		4	Gumbel-max	5.62%
			Green/red list	3.68%
			DiPMark	4.15%
			Ours	3.84%

Table 5: Same as Table **??**, but for the Llama model.

Model	Data	k	Metric	Gumbel-max	Ours
Llama-3.2-1B-Instruct	FinQA	2	Avg. Rejection Rate	22.28%	16.00%
			TPR	0.2300	0.1900
		4	Avg. Rejection Rate	22.83%	16.85%
			TPR	0.2100	0.2000
	ELI5	2	Avg. Rejection Rate	26.69%	18.42%
			TPR	0.3650	0.2800
		4	Avg. Rejection Rate	28.23%	19.42%
			TPR	0.4100	0.3600

Table 6: Same as Table **??**, but for the Llama model.

Model	Data	k/Num. Tokens	Metric	Max	Sum	HC$^+$
Llama-3.2-1B-Instruct	FinQA	2/382.170	TPR	0.335	**0.95**	0.905
			TPR aug.	0.125	**0.855**	0.665
		4/411.095	TPR	0.26	**0.975**	0.93
			TPR aug.	0.075	**0.84**	0.615
	ELI5	2/383.360	TPR	0.4	**1**	0.99
			TPR aug.	0.165	**0.965**	0.87
		4/414.440	TPR	0.385	**1**	**1**
			TPR aug.	0.12	**0.99**	0.905

Table 7: Comparison of the watermarking scheme with a single list and the watermarking scheme with disparate lists. The true positive rate (TPR) and the true positive rate after the substitution attack (TPR aug.) are reported in the form of (Sum statistic/HC$^+$ statistic).

Model	Data	k	Metric		Our Method single list	Our Method disparate lists
Phi-3-mini-4k-instruct (3.8B)	FinQA	2	S-BERT	mean	**0.8121**	0.7963
				median	**0.8431**	0.8382
			TPR		0.955/0.885	**0.975**/**0.92**
			TPR aug.		**0.905**/**0.73**	0.87/0.71
		4	S-BERT	mean	**0.808**	0.8023
				median	**0.8494**	0.8288
			TPR		**0.955**/0.89	0.95/**0.91**
			TPR aug.		**0.84**/**0.665**	0.83/0.625
	ELI5	2	S-BERT	mean	**0.7224**	0.7161
				median	**0.7499**	0.729
			TPR		**0.995**/**0.99**	0.99/0.98
			TPR aug.		**0.96**/**0.895**	0.935/0.865
		4	S-BERT	mean	**0.7237**	0.7185
				median	**0.7409**	0.7344
			TPR		**1**/**0.995**	**1**/0.99
			TPR aug.		**0.965**/0.82	0.955/**0.83**
Llama-3.2-1B-Instruct	FinQA	2	S-BERT	mean	0.8013	**0.8136**
				median	0.8314	**0.8347**
			TPR		**0.98**/**0.94**	0.95/0.905
			TPR aug.		**0.9**/**0.74**	0.855/0.665
		4	S-BERT	mean	**0.8128**	0.8117
				median	0.8402	**0.8426**
			TPR		0.97/0.925	**0.975**/**0.93**
			TPR aug.		**0.86**/**0.645**	0.84/0.615
	ELI5	2	S-BERT	mean	0.7287	**0.7351**
				median	0.7355	**0.7429**
			TPR		**1**/**0.99**	**1**/**0.99**
			TPR aug.		0.96/0.85	**0.965**/**0.87**
		4	S-BERT	mean	0.7243	**0.7315**
				median	**0.7453**	0.7433
			TPR		**1**/0.99	**1**/**1**
			TPR aug.		0.985/0.835	**0.99**/**0.905**

1.6.4 A Single Green List

Notice that our theory requires that the green lists $\mathcal{G}$ and the uniform random variables ζ be independent, and our scheme does not necessitate creating a new green list before generating each token. Instead, as long as the user does not have access to the watermark key, a priori, we can simply generate a single green list at the beginning and use it for generating all the tokens. This way, we only need to share a single pseudorandom function

to generate random variable ζ's. We performed an ablation study on this in Table 7. In particular, we compare the results of our watermarking scheme when we use a single green list and when we use disparate green lists. We report the BERT similarity scores and the TPR of the test statistic based on the sum of the ζ's and the higher criticism statistic (in the form TPR for sum/TPR for higher criticism). We conclude that having a single green list does not significantly impact our watermarking scheme when a larger context window is used to generate ζ. An analogous observation was also made in Zhao et al. [2023] for the green/red list soft watermarking scheme. However, it is unclear whether having the same green list for all the tokens will allow an adversary to have more power to learn the exact green list generator and produce attacks that can fool the detector, which is an interesting direction for future work.

References

S. Aaronson and H. Kirchner. Watermarking gpt outputs, 2023.

M. Abdin, S. A. Jacobs, A. A. Awan, J. Aneja, A. Awadallah, H. Awadalla, N. Bach, A. Bahree, A. Bakhtiari, H. Behl, et al. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*, 2024.

Z. Cai, S. Liu, H. Wang, H. Zhong, and X. Li. Towards better statistical understanding of watermarking llms. *arXiv preprint arXiv:2403.13027*, 2024.

C. Chen, S. Borgeaud, G. Irving, J.-B. Lespiau, L. Sifre, and J. Jumper. Accelerating large language model decoding with speculative sampling. *arXiv preprint arXiv:2302.01318*, 2023.

M. Christ, S. Gunn, and O. Zamir. Undetectable watermarks for language models. In *The*

Thirty Seventh Annual Conference on Learning Theory, pages 1125–1139. PMLR, 2024.

S. Dathathri, A. See, S. Ghaisas, P.-S. Huang, R. McAdam, J. Welbl, V. Bachani, A. Kaskasoli, R. Stanforth, T. Matejovicova, et al. Scalable watermarking for identifying large language model outputs. *Nature*, 634(8035):818–823, 2024.

D. Donoho and J. Jin. Higher criticism for detecting sparse heterogeneous mixtures1. *The Annals of Statistics*, 32(3):962–994, 2004.

A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Yang, A. Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

P. Fernandez, A. Chaffin, K. Tit, V. Chappelier, and T. Furon. Three bricks to consolidate watermarks for large language models. *arXiv preprint arXiv:2308.00113*, 2023.

E. J. Gumbel. *Statistical theory of extreme values and some practical applications: a series of lectures*, volume 33. US Government Printing Office, 1948.

Z. Hu, L. Chen, X. Wu, Y. Wu, H. Zhang, and H. Huang. Unbiased watermark for large language models. *arXiv preprint arXiv:2310.10669*, 2023.

B. Huang, B. Zhu, H. Zhu, J. D. Lee, J. Jiao, and M. I. Jordan. Towards optimal statistical watermarking. *arXiv preprint arXiv:2312.07930*, 2023.

J. Kirchenbauer, J. Geiping, Y. Wen, J. Katz, I. Miers, and T. Goldstein. A watermark for large language models. In *International Conference on Machine Learning*, pages 17061–17084. PMLR, 2023a.

J. Kirchenbauer, J. Geiping, Y. Wen, M. Shu, K. Saifullah, K. Kong, K. Fernando, A. Saha,

M. Goldblum, and T. Goldstein. On the reliability of watermarks for large language models. In *The Twelfth International Conference on Learning Representations*, 2023b.

R. Kuditipudi, J. Thickstun, T. Hashimoto, and P. Liang. Robust distortion-free watermarks for language models. *arXiv preprint arXiv:2307.15593*, 2023.

Y. Leviathan, M. Kalman, and Y. Matias. Fast inference from transformers via speculative decoding. In *International Conference on Machine Learning*, pages 19274–19286. PMLR, 2023.

X. Li, F. Ruan, H. Wang, Q. Long, and W. J. Su. A statistical framework of watermarks for large language models: Pivot, detection efficiency and optimal rules. *The Annals of Statistics (to appear)*, 2024.

J. Piet, C. Sitawarin, V. Fang, N. Mu, and D. Wagner. Mark my words: Analyzing and evaluating language model watermarks. *arXiv preprint arXiv:2312.00273*, 2023.

B. Saha and C. Ye. The i/o complexity of attention, or how optimal is flash attention? *arXiv preprint arXiv:2402.07443*, 2024.

S. Tu, Y. Sun, Y. Bai, J. Yu, L. Hou, and J. Li. Waterbench: Towards holistic evaluation of watermarks for large language models. *arXiv preprint arXiv:2311.07138*, 2023.

Y. Wu, Z. Hu, H. Zhang, and H. Huang. Dipmark: A stealthy, efficient and resilient watermark for large language models. *arXiv preprint arXiv:2310.07710*, 2023.

H. Zhang, B. L. Edelman, D. Francati, D. Venturi, G. Ateniese, and B. Barak. Watermarks in the sand: Impossibility of strong watermarking for generative models. *arXiv preprint arXiv:2311.04378*, 2023.

X. Zhao, P. Ananth, L. Li, and Y.-X. Wang. Provable robust watermarking for ai-generated text. *arXiv preprint arXiv:2306.17439*, 2023.

X. Zhao, L. Li, and Y.-X. Wang. Permute-and-flip: An optimally robust and watermarkable decoder for llms. *arXiv preprint arXiv:2402.05864*, 2024.