WATERMARKS FOR LANGUAGE MODEL VIA PROBABILISTIC AUTOMATA

Anonymous authors
Paper under double-blind review

ABSTRACT

A recent watermarking scheme for language models achieves distortion-free embedding and robustness to edit-distance attacks. However, it suffers from limited generation diversity and high detection overhead. In parallel, recent research has focused on undetectability—a property ensuring that watermarks remain difficult for adversaries to detect and spoof. In this work, we introduce a new class of watermarking schemes constructed through *probabilistic automata*. We present two instantiations: (i) a practical scheme with exponential generation diversity and computational efficiency, and (ii) a theoretical construction with formal undetectability guarantees under cryptographic assumptions. Extensive experiments on LLaMA-3B and Mistral-7B validate the superior performance of our scheme in terms of robustness and efficiency.

1 INTRODUCTION

The rapid development of large-scale language models (LMs) has markedly improved AI's ability to generate textual content (Brown et al., 2020). Despite these advancements, apprehensions have arisen over authenticity, ownership, and potential misuse of such technologies (Zellers et al., 2019; Solaiman et al., 2019). Traditional AI detection methods, such as classifier-based detection, often fall short in terms of robustness. In contrast, text watermarking offers a potential solution to these problems. It works by embedding a private key within the text that can be detected by the key holder, thereby identifying and minimizing the abuse of AI-generated content.

A widely adopted watermarking method conditions the decoder on the preceding k generated tokens (Kirchenbauer et al., 2023; Aaronson, 2022). While effective, this approach can degrade LM's output quality by introducing noticeable distortions, such as biases toward certain k-grams. To address these distortions, distortion-free watermarking was introduced to preserve the LM's output distribution (Kuditipudi et al., 2024); however, it does not guarantee LM's generation diversity. More recently, undetectable watermarking has been explored (Christ et al., 2024), which prevents detection by adversaries and naturally maintains generation diversity. Despite these advancements, the relationship between distortion-freeness and undetectability has thus far never been clearly defined. To this end, we establish the connection between distortion-freeness and undetectability, and show that many existing watermarks are detectable. One interpretation of this fact is that the watermarking output distribution can be recognized by probabilistic deterministic finite automata (PDFA) and can therefore be learned under the Probably Approximately Correct (PAC) framework.

Our work is closely related to the state-of-the-art watermarking approach introduced by Kuditipudi et al. (2024), which uses a cyclic key sequence as noise for unbiased decoding and leverages the *edit distance* (specifically, Levenshtein distance) metric to improve robustness against any edit-based attacks. However, this method suffers from two notable drawbacks: (1) it reduces generative diversity, often leading to deterministic outputs, and (2) it requires partitioning text into blocks with the time complexity scales quadratically with the block size, which creates a major computational bottleneck.

We introduce a new class of watermarking schemes represented by *probabilistic automata* (PA), with the following key contributions:

- Our framework generalizes the cyclic key sequence watermarking of Kuditipudi et al. (2024) as a special case, which can be modeled as a probabilistic deterministic finite automaton (PDFA) with a simple cyclic topology (see Figure 1).

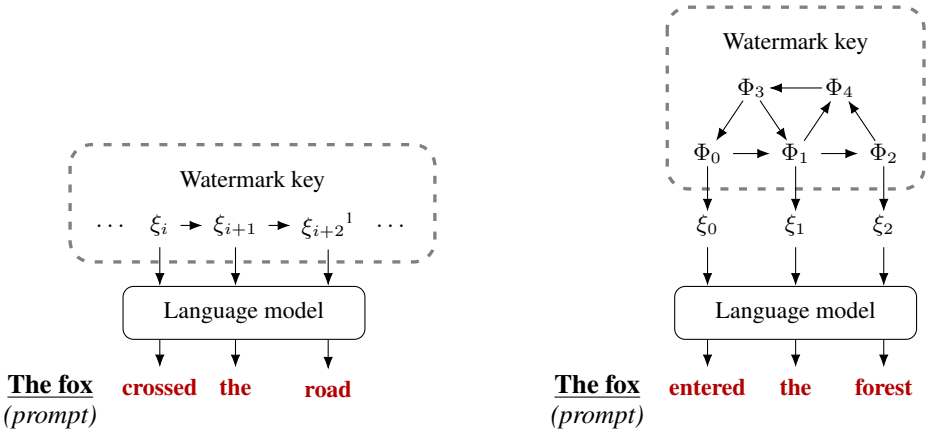

Figure 1: Comparison of generation of two watermarking schemes where the key follows a cyclic structure (Kuditipudi et al., 2024) on the left and a probabilistic automaton on the right. Each Φ specifies a probability distribution over ξ, with precise definitions provided in later sections.

- We extend this formulation to *probabilistic non-deterministic finite automata* (PNFA), a strictly more expressive class than PDFAs. Leveraging the fact that the class of languages recognized by PNFAs is not PAC-learnable under the sparse Learning Parity with Noise (LPN) assumption (Kearns et al., 1994), we construct an undetectable watermarking scheme.
- We instantiate a practical scheme that significantly improves both generation diversity and detection efficiency over Kuditipudi et al. (2024): (i) Increasing generation diversity from $\Theta(\lambda)$ to $\Omega(\lambda d^n)$, where $d \geq 1$, λ is the key length, and n is the sequence length, and (ii) Reducing detection time complexity from $\Theta(\lambda n k^2)$ to $\Theta(\lambda n)$, where k is the block size.

2 Related Work

Watermarking of language models. Text watermarking aims at blending a private "key" in text generation so that it can be detected by key holders. Early approaches relied on subtle text modifications using heuristics Atallah et al. (2001; 2002); Topkara et al. (2005). Following the autoregressive nature of language models (LMs), recent watermarking methods start to condition the token generation by the key and k prior tokens (Kirchenbauer et al., 2023; Aaronson, 2022; Zhao et al., 2024). However, these methods can significantly alter the underlying LM's distribution, for instance, by introducing biases for certain k-grams.

Watermarking without changing the next token distribution of the LM on a *single text sample* is defined as *distortion-free* watermarking (Kuditipudi et al., 2024; Hu et al., 2024). For example, Kuditipudi et al. (2024) uses a cyclic key sequence of noise for unbiased decoding and was the first to employ the *edit distance* metric as the alignment between text and the key sequence for detection. While this method enhances the robustness of watermarking against edit-based attacks, it suffers from two drawbacks of lacking generation diversity, and it relies on partitioning the text into blocks and repeatedly shifting the key sequence to compute edit distances multiple times, posing significant efficiency challenges.

Recent works have constructed *undetectable* watermarks theoretically that require watermarked texts to be indistinguishable across *multiple queries*, yet none of them remains practical. For instance, Christ et al. (2024)'s construction is based on hash functions and not robust to edit-based attacks. Christ & Gunn (2024) assume a binary symmetric channel model for LMs, which is clearly unrealistic. Golowich & Moitra (2024) make an assumption that the vocabulary size scales polynomially with the security parameter (i.e., the size of the "key"), which does not typically hold in practice. Notably, all of these undetectable watermarking schemes rely on the construction of pseudorandom functions, yet they are based on disparate assumptions and lack a unified framework.

Probabilistic automata. Probabilistic automata (PA) are widely studied in computational linguistics that describe distributions with latent variables over finite sequences of symbols. The class of PA consists of probabilistic nondeterministic finite automata (PNFA) and their proper subclass, proba-

[1] All subscripts are taken modulo λ.

bilistic deterministic finite automata (PDFA). The learnability of these automata has gained profound theoretical interest and practical relevance, particularly in modeling distributions over strings.

Kearns et al. (1994) explored the complexity of learning PDFAs within the Probably Approximately Correct (PAC) framework. Specifically, they established that PAC-learning PDFAs with a two-letter vocabulary is at least as hard as PAC-learning noisy parities, which is believed to be computationally hard. Consequently, the entire class of PDFAs cannot be PAC-learned within polynomial time constraints. On the other hand, Ron et al. (1995); Clark & Thollard (2004) showed that under the constraints of a certain distinguishability on the states, acyclic PDFA and cyclic PDFAs with bounded expected string length from any states are PAC-learnable.

Although there have been works on learning PDFAs and intermediate forms of PNFAs, few studies address the problem of learning general PNFAs due to their hardness. Terwijn (2002) demonstrated that PNFAs are not PAC-learnable if Blum integer factorization is hard. Angluin & Kharitonov (1991) showed that learning remains hard even if adversaries have oracle access to membership queries (in the context of watermarking the queries of the detection function).

Edit distance and error correction. Edit distance quantifies the similarity between two strings by counting the minimum number of operations required to transform one into the other. This concept extends to formal languages, where the edit distance is defined as the minimum distance between any pair of sequences within the languages. This metric is particularly valuable in error correction, where it helps identify the closest valid string to a given input string. Wagner (1974) introduced an error-correcting algorithm that constructs a finite state automaton to recognize a set of strings. Their approach uses dynamic programming to compute the edit distance between a string and a regular language.

3 PRELIMINARIES

We denote the alphabet of an automaton by Σ and the vocabulary of a language model by $\mathcal{V}$ to explicitly distinguish them. A language is defined as a mapping $\Sigma^* \to [0,1]$ (or $\mathcal{V}^* \to [0,1]$, respectively). Given an automaton $\mathcal{M}$, we denote by $\mathcal{L}(\mathcal{M}) \subseteq \Sigma^*$ the language recognized by $\mathcal{M}$, i.e., the set of strings accepted by $\mathcal{M}$. We denote the size of the watermark key by λ. In undetectable watermarking settings, it also serves as the security parameter, which measures the strength of distortion-freeness and undetectability of a watermarking scheme. Further discussion on λ will follow in Section 5. The length of the generated sequence is given by $m = \mathcal{O}(\text{poly}(\lambda))$, where $\text{poly}(\cdot)$ denotes a polynomial function. For a sequence $\boldsymbol{x} = (x_1, \ldots, x_n)$, we write $\boldsymbol{x}_{i:} = (x_i, \ldots, x_n)$ for a suffix, and $\boldsymbol{x}_{i:j} = (x_i, \ldots, x_j)$ for a contiguous subsequence.

Definition 1 (Language Model)**.** An (autoregressive) language model is defined by a function $\mathsf{Model} : \mathcal{V}^* \to \Delta(\mathcal{V})$ that maps a sequence of tokens to a probability distribution over $\mathcal{V}$, where $\Delta(\cdot)$ denotes a probability distribution over a set. Given an initial sequence of tokens (a *prompt*) $\boldsymbol{x} \in \mathcal{V}^*$, the probability of a sequence $\boldsymbol{y} = (y_1, y_2, \ldots, y_m)$ is defined as

$$p(\boldsymbol{y}) = \prod_{i=1}^{m} p(y_i \mid \boldsymbol{x}, \boldsymbol{y}_{1:i-1}), \tag{1}$$

where $\boldsymbol{y}_{l:h} = (y_l, y_{l+1}, \ldots, y_h)$ represents a subsequence of $\boldsymbol{y}$. Each conditional probability is modeled by

$$p(\cdot \mid \boldsymbol{x}, \boldsymbol{y}_{1:i-1}) = \mathsf{Model}(\boldsymbol{x}, \boldsymbol{y}_{1:i-1}). \tag{2}$$

We use the notation $\boldsymbol{y} \xleftarrow{\mathsf{AR}} \mathsf{Model}(\boldsymbol{x})$ to indicate that $\boldsymbol{y}$ is autoregressively generated by Model given $\boldsymbol{x}$, following Equation 1.

Definition 2 (Decoder-based Watermarking Scheme)**.** A decoder-based watermarking scheme is a triplet $\mathcal{W} := (\mathsf{Gen}, \mathsf{Model}^{\mathsf{wat}}, \mathsf{Detect})$ such that:

1. The key generation algorithm Gen is randomized and takes as input 1^λ to generate a secret key $\mathsf{sk} \in \mathcal{K}$: [2]

$$\mathsf{sk} \leftarrow \mathsf{Gen}(1^\lambda). \tag{3}$$

[2]The input is given in unary notation to ensure polynomial runtime in λ. The structure of the key can be arbitrary, with specifics described in later sections.

2. The watermarking algorithm $\mathsf{Model}_{\mathsf{sk}}^{\mathsf{wat}} := (\mathsf{Model}, \Phi, \Gamma)$ is an autoregressive process that consists of an unwatermarked model Model, a noise generator $\Phi : \mathcal{K} \times \Xi^* \times \mathcal{V}^* \to \Delta(\Xi)$, and a decoder $\Gamma : \Xi \times \Delta(\mathcal{V}) \to \mathcal{V}$, where Ξ is the noise space. At the i-th decoding step, $\xi_i \in \Xi$ is sampled by

$$\xi_i \sim \Phi_{\mathsf{sk}}(\boldsymbol{\xi}_{1:i-1}, \boldsymbol{x}, \boldsymbol{y}_{1:i-1}), \tag{4}$$

then the next token y_i is produced by the decoder deterministically given the noise ξ_i:

$$y_i \leftarrow \Gamma(\xi_i, \mathsf{Model}(\boldsymbol{x}, \boldsymbol{y}_{1:i-1})). \tag{5}$$

3. The detection algorithm $\mathsf{Detect}_{\mathsf{sk}}$ takes as input sk and $\boldsymbol{y}$, and outputs true or false.

While the above definition presented seems abstract, it covers a wide range of existing watermarking frameworks. Specific examples are provided in Appendix E.

Ideally, $\mathsf{Detect}_{\mathsf{sk}}(\boldsymbol{y})$ should output true if $\boldsymbol{y}$ is generated by $\mathsf{Model}_{\mathsf{sk}}^{\mathsf{wat}}(\boldsymbol{x})$ for some $\boldsymbol{x}$, and output false if $\boldsymbol{y}$ is independent of sk. The former property is referred to as *completeness* and the latter *soundness*.

As a special case, when Φ does not depend on the prefix of tokens $\boldsymbol{x}$ and $\boldsymbol{y}_{1:i-1}$, Equation 4 simplifies to

$$\xi_i \sim \Phi_{\mathsf{sk}}(\boldsymbol{\xi}_{1:i-1}). \tag{6}$$

We refer to this case as *model-agnostic* as the distribution of the noise can be decomposed autoregressively with the chain rule and does not depend on the specific model used and therefore can be precomputed before decoding.

All decoder-based watermarking schemes require sufficiently high text entropy; otherwise, the output tends to be deterministic and no watermarks can be embedded. For a detailed discussion of entropy effects, refer to Christ et al. (2024); Kuditipudi et al. (2024); we do not repeat these efforts here. In the sequel a watermarking scheme always denotes a decoder-based watermarking scheme.

The robustness of a watermarking scheme quantifies its ability to withstand edit-based corruptions to the watermarked data without losing the embedded watermark.

Definition 3 (Edit Distance)**.** The edit distance $d(\boldsymbol{s}_1, \boldsymbol{s}_2)$ between two sequences $\boldsymbol{s}_1, \boldsymbol{s}_2 \in \mathcal{V}^*$ is the minimum cost of transforming $\boldsymbol{s}_1$ into $\boldsymbol{s}_2$ through a sequence of single-position edit operations, including insertion, deletion, and substitution.

Definition 4 (Robustness)**.** A watermarking scheme is considered robust if, for any watermarked sequence $\boldsymbol{y}$ from $\mathsf{Model}_{\mathsf{sk}}^{\mathsf{wat}}(\boldsymbol{x})$ and any sequence $\boldsymbol{y}'$ with the edit distance bounded by $d(\boldsymbol{y}, \boldsymbol{y}') \leq \gamma \max(|\boldsymbol{y}|, |\boldsymbol{y}'|)$ for some $\gamma > 0$, the detection function reliably identifies the watermark:

$$\mathsf{Detect}_{\mathsf{sk}}(\boldsymbol{y}') = \mathsf{true}. \tag{7}$$

Note that a sequence detected as watermarked is not necessarily generated by the watermarked model. Robustness ensures that a watermarked sequence and its close neighbors remain detectable, but this does not compromise the property of soundness.

4 Constructing watermarks through probabilistic automata

We begin by introducing the relevant definitions that underlie our watermarking constructions.

Definition 5 (Probabilistic Non-Deterministic Finite Automaton)**.** A probabilistic non-deterministic finite automaton (PNFA) defined as a tuple $(Q, \Sigma, \delta, \pi_0, \pi_f)$, where (1) Q is a finite set of states; (2) Σ is a finite alphabet of input symbols; (3) $\delta : Q \times \Sigma \times Q \to [0, 1]$ is the transition probability function; (4) $\pi_0 : Q \to [0, 1]$ defines the initial probability of each state; (5) $\pi_f : Q \to [0, 1]$ defines the final probability of each state.

Definition 6 (Probabilistic Deterministic Finite Automaton)**.** A probabilistic non-deterministic finite automaton (PNFA) $(Q, \Sigma, \delta, \pi_0, \pi_f)$ is a probabilistic deterministic finite automaton (PDFA) if: (1) $\exists q_0 \in Q$ such that $\pi_0(q_0) = 1$ and $\forall q \in Q \setminus \{q_0\}$, $\pi_0(q) = 0$, and (2) $\forall q \in Q, \forall a \in \Sigma$, there exists at most one state $q' \in Q$ such that $\delta(q, a, q') > 0$.

4.1 WATERMARKING SCHEMES REPRESENTED BY PROBABILISTIC AUTOMATA

In model-agnostic watermarking schemes, given a secret key sk, Φ_{sk} defines a distribution of the random variable $\boldsymbol{\xi} \in \Xi^*$ by applying Equation 6 autoregressively. One case of this distribution is the stochastic language recognized by a PA. We model this distribution using a hierarchical automaton. Specifically, for a PA $\mathcal{M} = (Q, \Sigma, \delta, \pi_0, \pi_f)$ with $\Sigma \subset \Delta(\Xi)$, the process starts from an initial state q_0. Each transition produces $\Phi_i \in \Sigma$, a probability distribution over Ξ, from which noise $\xi_i \sim \Phi_i$ is sampled. The noise ξ_i, represented by a binary sequence, is used for decoding the next token y_i as described in Equation 5. Each probability distribution Φ_i is modeled by a subordinate PA with a binary alphabet. The hierarchical structure allows the PA to model the noise distribution represented by binary alphabet.

4.2 CONSTRUCTING WATERMARKS FOR LANGUAGE MODELS

We now elaborate a specific construction of the watermarking scheme. We consider a decoder that uses exponential minimum sampling following Aaronson (2022); Kuditipudi et al. (2024). The decoder generates the next token based on a noise ξ and the model's output probabilities, which can be formally expressed as

$$\Gamma(\xi_i, \mathsf{Model}(\boldsymbol{y}_{1:i-1})) = \arg\min_{j \in \mathcal{V}} (\pi_j / \log(\mu_j)), \tag{8}$$

where π_j is the probability assigned by $\mathsf{Model}(\boldsymbol{y}_{1:i-1})$ to token $j \in \mathcal{V}$, and $\xi_i = (\mu_1, \ldots, \mu_{|\mathcal{V}|})$ with $\mu_j \overset{\text{i.i.d.}}{\sim} \text{Uniform}[0, 1]$. This decoder preserves the model's categorical distribution at each step.

Transitioning from continuous to binary representation, any real number $z \in [0, 1)$ can be approximated using its binary expansion:

$$z = \frac{1}{2^c} \sum_{i=0}^{c-1} 2^i \sigma_i, \quad \sigma_i \in \{0, 1\}, \tag{9}$$

where c denotes the precision level, and σ_i represents the i-th bit of the binary expansion of z.

We introduce a PA to generate a sequence of noise. The PA consists of λ virtual states, each corresponding to a subordinate PA that models a specific noise distribution. These states are labeled as $q_0, q_1, \ldots, q_{\lambda-1}$ with each state q_i transitioning to $q_{i+1 \bmod \lambda}, \ldots, q_{i+d \bmod \lambda}$ with equal probability, thereby forming a d-regular graph. The arrangement of these states is strategically designed for robustness and efficiency.

The subordinate probabilistic automaton is defined over a vocabulary of size $|\mathcal{V}|$, with bitwidth b and $c \geq b$. At each decoding step, it generates a binary noise vector $\xi = (\mu_1, \ldots, \mu_{|\mathcal{V}|})$, where each $\mu_i \in \{0, 1\}^c$.

The automaton begins at an initial state q_0 and terminates at a final state q_f, progressing through $|\mathcal{V}|$ layers that each encode a binary vector μ_i. The first layer starts with: $q_0 \to \sigma_{1,1}$, and each layer proceeds through intermediate bitwise states: $\sigma_{i,j} \to \sigma_{i,j+1}$, for $1 \leq i \leq |\mathcal{V}|$, $1 \leq j < b$, where $\sigma_{i,j}$ encodes the j-th bit of μ_i. At $\sigma_{i,b}$, the automaton branches into two parallel Boolean paths: $\sigma_{i,b} \to \iota_{i,b+1}, \sigma_{i,b} \to \hat{\iota}_{i,b+1}$, which continue as: $\iota_{i,j} \to \iota_{i,j+1}$, $\iota_{i,j} \to \hat{\iota}_{i,j+1}$, $\hat{\iota}_{i,j} \to \iota_{i,j+1}$, $\hat{\iota}_{i,j} \to \hat{\iota}_{i,j+1}$, where $\iota_{i,j} = 0$ and $\hat{\iota}_{i,j} = 1$ represent bit encodings of μ_i. Between layers, transitions connect the terminal states of layer i to the initial states of layer $i{+}1$: $\iota_{i,c} \to \sigma_{i+1,1}, \hat{\iota}_{i,c} \to \sigma_{i+1,1}$, for $b \leq j < c$, and the automaton concludes after the final layer with $\iota_{|\mathcal{V}|,c}, \hat{\iota}_{|\mathcal{V}|,c} \to q_f$.

For each state, all outgoing transitions have equal probability. As a special case where $d = 1$ and sufficiently large b and c, the PA produces noise equivalent to the cyclic key sequence watermarking (Kuditipudi et al., 2024). The PA sampling process is integrated into the token generation, as described in Algorithm 1. An illustration of a subordinate PA is provided in Appendix H.1.

We now proceed to describe the detection algorithm. We begin by defining a cost following Aaronson (2022) and Kuditipudi et al. (2024) as

$$d_0(y, \xi) = \log(1 - \mu_y), \tag{10}$$

where $\xi = (\mu_1, \mu_2, \ldots, \mu_{|\mathcal{V}|})$ and $\mu_i \in [0, 1]$ is represented by its binary expansion using the subordinate PA. For $\Phi \in \Delta(\Xi)$, the cost is defined as

$$d_0(y, \Phi) = \max_{\xi \in \text{supp}(\Phi)} \{d_0(y, \xi)\}. \tag{11}$$

Before we define the Levenshtein distance for PAs, we introduce the necessary definition.

Definition 7 (Support Automaton)**.** The support automaton of a PA $\mathcal{M} = (Q, \Sigma, \delta, \pi_0, \pi_f)$ is a non-deterministic finite automaton (NFA) $\underline{\mathcal{M}} = (Q, \Sigma, \delta, Q_0, Q_f)$, where Q_0 (respectively Q_f) denotes the set of initial (respectively final) states, and $\delta \subseteq Q \times \Sigma \times Q$ is the transition function defined as $(q, a, q') \in \delta \Leftrightarrow \phi(q, a, q') > 0$.

Algorithm 1 Watermarking Schemes Represented by PAs

Input: Model, PA $\mathcal{M}$, decoder Γ
Output Watermarked sequence $\boldsymbol{y}$
1: Sample an initial state q from $\mathcal{M}$.
2: **for** each decoding step i **do**
3: | Transition q to the next state and obtain $\Phi_i \in \Delta(\Xi)$.
4: | Sample $\xi_i \sim \Phi_i$.
5: | Select the next token y_i using ξ_i via Equation 8.
6: **return** $\boldsymbol{y}$

Definition 8 (Generalized Levenshtein Distance)**.** For a sequence $\boldsymbol{y} \in \mathcal{V}^*, \boldsymbol{\Phi} \in \Delta^*(\Xi), \gamma_i, \gamma_d$, the Levenshtein distance between $\boldsymbol{y}$ and $\boldsymbol{\Phi}$ is defined recursively as

$$d_L(\boldsymbol{y}, \boldsymbol{\Phi}) = \min \begin{cases} \gamma_d|\boldsymbol{y}| + \gamma_i|\boldsymbol{\Phi}|, \text{if } |\boldsymbol{y}| = 0 \text{ or } |\boldsymbol{\Phi}| = 0, \\ d_L(\boldsymbol{y}_{2:}, \boldsymbol{\Phi}) + \gamma_d, \\ d_L(\boldsymbol{y}, \boldsymbol{\Phi}_{2:}) + \gamma_i, \\ d_L(\boldsymbol{y}_{2:}, \boldsymbol{\Phi}_{2:}) + d_0(y_1, \Phi_1). \end{cases} \tag{12}$$

For a PA $\mathcal{M}$ defining a support language $\mathcal{L}(\underline{\mathcal{M}}) \subseteq \Delta^*(\Xi)$, the Levenshtein distance between $\boldsymbol{y}$ and $\mathcal{M}$ is defined as

$$d_L(\boldsymbol{y}, \mathcal{M}) = \min_{\boldsymbol{\Phi} \in \mathcal{L}(\underline{\mathcal{M}})} \{d_L(\boldsymbol{y}, \boldsymbol{\Phi})\}. \tag{13}$$

In Equation 12, the behavior of $\boldsymbol{y}$ and $\boldsymbol{\Phi}$ is not symmetric, as the length of $\boldsymbol{\Phi}$ can be infinite. Therefore, we assign different costs to insertion (γ_i) and deletion (γ_d). Meanwhile, the generalized Levenshtein distance is different from the design in Kuditipudi et al. (2024). In their algorithm, the lengths of the key sequence and the text are constrained to be equal, and it may disrupt their alignment when insertions or deletions are involved. In contrast, our definition allows flexible alignment despite difference in lengths.

The naive dynamic programming computes Equation 13 in $\mathcal{O}(dm\lambda \log \lambda))$ time (Wagner, 1974), whereas our PNFA construction reduces it to $\mathcal{O}(m\lambda)$. The dynamic programming algorithm is detailed in Appendix D.1.

Empirical p-value. We quantify the alignment between an observed sequence $\boldsymbol{y}$ and the watermarked model $\mathcal{M}_{\mathsf{sk}}$ by the statistic $\psi = d_L\big(\boldsymbol{y}, \mathcal{M}_{\mathsf{sk}}\big)$, and treat as null hypothesis H_0 that $\boldsymbol{y}$ is independent of sk. We sample N independent keys $\{\mathsf{sk}_i\}_{i=1}^N$ and compute $\psi_i = d_L\big(\boldsymbol{y}, \mathcal{M}_{\mathsf{sk}_i}\big) \quad (i = 1, \ldots, N)$. The empirical p-value is then

$$\hat{p} = \frac{1 + \sum_{i=1}^N \mathbf{1}\big[\psi_i \leq \psi\big]}{N + 1}. \tag{14}$$

A small p-value indicates that $\boldsymbol{y}$ aligns unusually well with the true key (suggesting a watermark), whereas a larger p-value is consistent with unwatermarked noise.

An upper bound of p-value via Vysochanskij–Petunin Inequality. Let μ and σ^2 denote the sample mean and variance of $\{\psi_i\}_{i=1}^N$, given by $\mu = \frac{1}{N}\sum_{i=1}^N \psi_i$ and $\sigma^2 = \frac{1}{N-1}\sum_{i=1}^N (\psi_i - \overline{\psi})^2$, and define the standardized statistic

$$z = (\psi - \mu)/\sigma. \tag{15}$$

Under the mild assumption that the null distribution of d_L is unimodal, the one-sided Vysochanskij–Petunin inequality (Mercadier & Strobel, 2021) gives the following upper bound:

$$\Pr[Z \leq z] \leq \begin{cases} 4/9(z^2 + 1) & \text{if } |z| \geq \sqrt{5/3}, \\ 4/3(z^2 + 1) - 1/3 & \text{otherwise.} \end{cases} \tag{16}$$

We refer to this scheme as WEPA, for Watermarking schemE through Probabilistic Automata. The design of WEPA makes it robust and highly efficient for dynamic programming. As observed, disregarding bitwidth constraints, the generation diversity of WEPA is $\Theta(\lambda d^m)$, which is a substantial improvement over the construction of Kuditipudi et al. (2024).

5 Expressiveness of PA-based watermarking schemes

In this section, we draw connections between distortion-freeness, undetectability, and security levels in watermarking schemes. The security level is measured by the security parameter λ that determines the key size and reflects the computational hardness of breaking cryptographic schemes. It is assumed to be known to adversaries, and their running time is expressed as functions of λ rather than concrete values. An adversary is considered efficient if it can be modeled as a probabilistic algorithm with running time bounded by a polynomial function of λ. A function is called *negligible* if it becomes asymptotically smaller than the inverse of any polynomial as λ grows, which represents a probability that is practically negligible for large λ. Negligible functions can be expressed as $\text{negl}(\lambda) = \mathcal{O}(\frac{1}{\text{poly}(\lambda)})$ (i.e., functions that vanish faster than any inverse polynomial).

Distortion-freeness. We begin the definition by considering an adversary who has oracle access to the unwatermarked model (i.e., can make unlimited queries) and observes a single instance of watermarked data, and its goal is to determine whether the observed data is watermarked or not. While we make assumptions about the adversary's computational capabilities, we make no assumptions about its strategy, which means the definition provides protection against any computationally bounded adversary. The idea is that the adversary should be unable to extract any partial information that distinguishes watermarked data from unwatermarked data.

In addition, we do not make assumptions about the adversary's prompting strategy. The adversary can manipulate the model's behavior by prompting so that the model outputs arbitrary distribution. This ensures that the security definition remains robust against adversarial control over the model's output distribution. Since the adversary can shape the output distribution arbitrarily, the specific input to the model becomes irrelevant—choosing a prompt equates to choosing a model. Therefore, we can safely omit the model's input in our analysis.

The definition is formalized using the indistinguishability against one-time attacks ($\mathsf{IND\text{-}OT}$) game, defined for a watermarking scheme $\mathcal{W}$, an adversary $\mathcal{A}$, and the security parameter λ: (1) A secret key is generated via $\mathsf{sk} \leftarrow \mathsf{Gen}(1^\lambda)$; (2) $\mathcal{A}$ is given an input 1^λ and chooses $\mathsf{Model}(\cdot)$; [3] (3) a uniform bit $b \in \{0, 1\}$ is selected. A sequence $\boldsymbol{y} \xleftarrow{\mathsf{AR}} f(\cdot)$ is computed and given to $\mathcal{A}$, where $f = \mathsf{Model}$ if $b = 0$ otherwise $f = \mathsf{Model}^{\mathsf{wat}}_{\mathsf{sk}}$; (4) finally, $\mathcal{A}$ outputs a bit b'. The advantage of $\mathcal{A}$ in this game is defined as

$$\text{Adv}^{\mathsf{IND\text{-}OT}}_{\mathcal{W}}(\mathcal{A}) := \left| \Pr\left[b = b'\right] - \frac{1}{2} \right|. \tag{17}$$

Definition 9 (Distortion-freeness)**.** A watermarking scheme $\mathcal{W}$ is *distortion-free* if it is indistinguishable against one-time attacks. Formally, for any polynomial-time $\mathcal{A}$,

$$\text{Adv}^{\mathsf{IND\text{-}OT}}_{\mathcal{W}}(\mathcal{A}) \leq \text{negl}(\lambda). \tag{18}$$

Moreover, a watermarking scheme is considered *perfectly distortion-free* if $\text{Adv}^{\mathsf{IND\text{-}OT}}_{\mathcal{W}}(\mathcal{A}) = 0$.

The following theorem provides an equivalent definition of perfect distortion-freeness.

Theorem 1. *A watermarking scheme is perfectly distortion-free iff for every language model that defines language distribution $\mathcal{L}$, every $\boldsymbol{y} \in \mathcal{L}$, it follows that*

$$\Pr[Y = \boldsymbol{y}] = \mathbb{E}_{\mathsf{sk} \leftarrow \mathsf{Gen}(1^\lambda)}\big[\Pr[Y = \boldsymbol{y} \mid K = \mathsf{sk}]\big], \tag{19}$$

where K is the random variable of the secret key.

We defer this proof to Appendix F.1. Theorem 1 indicates that the randomness in generating an output from the unwatermarked model can be equivalently represented by first sampling a key from the key space and then generating the sequence conditioned on that key. Consequently, the original distribution of the model's outputs is preserved. However, this condition is often too strict, and even

[3] We omit the input $\boldsymbol{x}$ to Model as choosing a prompt equates to choosing a model as explained.

Kuditipudi et al. (2024)'s construction is not perfectly distortion-free if the text length is greater than λ. Instead, our definition emphasizes *computational indistinguishability*, which is more attainable and aligns with the limitations of real-world systems.

Undetectability. Distortion-freeness ensures that the quality of the language model is not changed by preserving the distribution in one query. It is nonetheless detectable with multiple queries, and a higher level of security is desired. Consider the following indistinguishability against chosen prompt attack (IND-CPA) game, defined for a watermarking scheme $\mathcal{W}$, an adversary $\mathcal{A}$, and the security parameter λ: (1) A secret key is generated via $\mathsf{sk} \leftarrow \mathsf{Gen}(1^\lambda)$; (2) $\mathcal{A}$ is given an input 1^λ and the oracle access to $\mathsf{Model}'^{\mathsf{wat}}_{\mathsf{sk}}(\cdot)$ for any Model', and chooses $\mathsf{Model}(\cdot)$; (3) a uniform bit $b \in \{0, 1\}$ is selected. A sequence $\boldsymbol{y} \xleftarrow{\mathsf{AR}} f(\cdot)$ is computed and given to $\mathcal{A}$, where $f = \mathsf{Model}$ if $b = 0$ otherwise $f = \mathsf{Model}^{\mathsf{wat}}_{\mathsf{sk}}$; (4) finally, $\mathcal{A}$ outputs a bit b'. The advantage of $\mathcal{A}$ in this game is defined as

$$\mathrm{Adv}^{\mathsf{IND\text{-}CPA}}_{\mathcal{W}}(\mathcal{A}) := \left| \Pr\left[b = b'\right] - \frac{1}{2} \right|. \tag{20}$$

Definition 10 (Undetectability)**.** A watermarking scheme $\mathcal{W}$ is *undetectable* if, for all polynomial-time adversaries, there exists a negligible function negl such that

$$\mathrm{Adv}^{\mathsf{IND\text{-}CPA}}_{\mathcal{W}}(\mathcal{A}) \leq \mathrm{negl}(\lambda). \tag{21}$$

By definition it is clear that all undetectable watermarking schemes are distortion-free due to the stronger capabilities of adversaries. This implies that undetectable watermarking maintains the output distribution of the watermarked model unchanged. Conversely, any watermarking scheme that introduces distortion is also detectable. This further implies that the watermarks in Kirchenbauer et al. (2023); Zhao et al. (2024); Aaronson (2022); Kuditipudi et al. (2024) are all detectable.

The following theorem highlights the *expressiveness* (i.e., the extent to which the automaton represents various sequences) of PAs for constructing watermarking schemes.

Theorem 2. *(Abridged) Under certain conditions, there exists an undetectable watermarking scheme that can be represented by a PA if sparse LPN is hard.*

See Appendix G for the full statement and proof. With this framework, the cyclic key sequence watermarking scheme (Kuditipudi et al., 2024) can be represented by a PDFA with λ states and cyclic transitions, though its expressiveness is limited by its simple structure. To achieve higher expressiveness, we consider using PNFAs, which are strictly more expressive than PDFAs.

6 Empirical evaluation

We conducted an empirical evaluation of WEPA's statistical properties and robustness on LLaMA-3.2-3B (Dubey et al., 2024) and Mistral-7B (v0.3) (Jiang et al., 2023) language models. In line with prior works (Kirchenbauer et al., 2023; Kuditipudi et al., 2024), we generated watermarked text continuations from the news-like subset of the C4 dataset (Raffel et al., 2020). We compared WEPA with three baselines: cyclic key sequence watermarking (Seq) with exponential minimum sampling (Kuditipudi et al., 2024), unigram green-red set watermarking (G/R) (Kirchenbauer et al., 2023), and unbiased watermarking (Unbiased) (Hu et al., 2024). Notably, the G/R baseline is not directly comparable, as it may introduce noticeable distortion. For WEPA and Seq, we computed p-values via Equation 14 with a sample size of 10,000 for consistency. For Unbiased, we report the upper bound of p-values as described in Hu et al. (2024). All methods were evaluated using their strongest hyperparameter settings as recommended by the original authors; details are provided in Appendix C.2. For each experiment, we report the $1/3$, $1/2$ (median), and $2/3$ quantiles of p-values for watermarked text across 100 samples. Further empirical results are included in Appendix B due to page limits.

6.1 Varying text lengths

We varied the generation length of watermarked text from 4 to 20 tokens, as shown in Figure 2. The results highlight differences between the two language models due to variations in their generation entropy. Among the methods, WEPA ($d = 1$) and Seq consistently achieve the strongest detection performance. WEPA ($d = 2$) performs slightly worse, followed by G/R. The Unbiased method lags behind, especially on shorter sequences. Full numerical results are provided in Appendix B.7, and

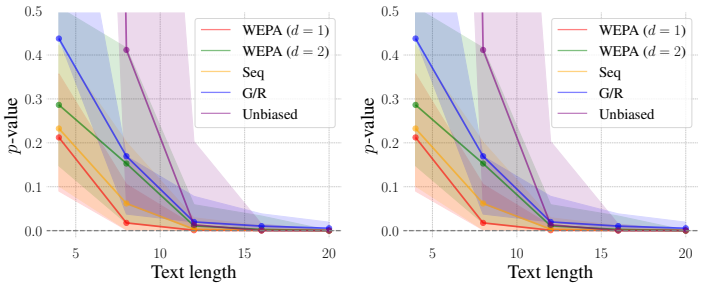

Figure 2: Median p-values varying text lengths on LLaMA-3B (left) and Mistral-7B (right).

Scheme	**Time** (s)
WEPA ($d = 1$)	7.81
WEPA ($d = 2$)	8.22
WEPA ($d = 1, b = 6$)	10.60
Seq (our impl.)	2039.56
Seq	49024.57

Table 1: Detection efficiency comparison. Unless otherwise specified, `float32` is used for b in WEPA.

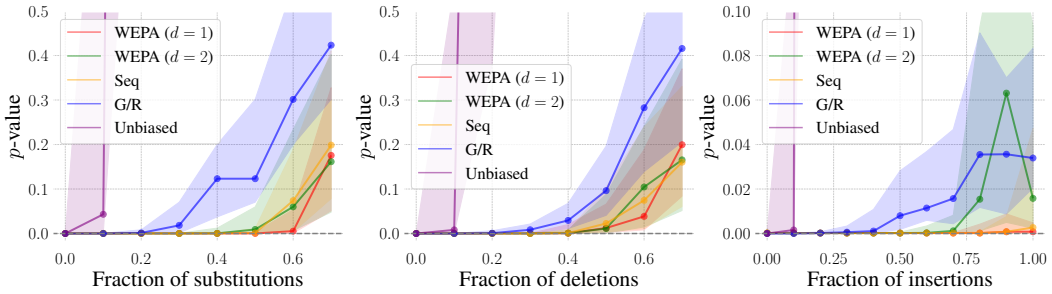

Figure 3: Median p-values under substitution (left), deletion (middle), and insertion (right) attacks.

we further report ROC-AUC and TPR@1%FPR in Appendix B.1. Notably, watermarking short text is inherently more challenging than long text. For completeness, we also include experiments on longer sequences (up to 200 tokens) in Appendix B.2.

6.2 Robustness to edit-based attacks

We evaluated the robustness against three types of edit-based attacks: (1) substitution, (2) deletion, and (3) insertion, each applied by randomly corrupting a fraction of tokens. For substitution and insertion, replacement tokens were sampled uniformly from the vocabulary. No padding was introduced, nor was truncation applied following deletions or insertions. All experiments were conducted on sequences of length 50 tokens.

Figure 3 reports results on LLaMA-3B, while we defer the results on Mistral-7B in Appendix B.6. WEPA ($d = 1$) performs slightly better than the other methods across all attack types, likely due to its flexible alignment mechanism. In contrast, Seq performs slightly worse, as it requires exact alignment between the generated text and a fixed-length key sequence. G/R shows reduced robustness, likely due to its unigram-based design. The Unbiased method is not robust under these attacks, as it was not designed to handle edit-based perturbations.

We note that our watermark is specifically designed to resist edit-based perturbations, and does not aim to defend against semantic or paraphrasing attacks. As such, we do not include experiments on paraphrasing, which fall outside the scope of our work.

6.3 Efficiency

We compared the runtime of the detection algorithms for WEPA ($d = 1$), WEPA ($d = 2$) and Seq. For WEPA ($d = 2, b = 6$), we choose $b = 6$ as the bitwidth does not affect efficiency. The results are presented in Table 1. For Seq, we implemented an optimization with token discretization. Each algorithm was evaluated on a sample from C4 dataset with text length of 256 using LLaMA's tokenizer. WEPA is significantly faster than Seq as predicted by the different time complexities.

7 Conclusion

We introduced a class of watermarking schemes constructed through probabilistic automata. Within this framework, we instantiated WEPA, a practical watermarking method that achieves both improved generation diversity and more efficient detection. Empirical results further demonstrate its effectiveness and efficiency. Furthermore, by extending to probabilistic non-deterministic finite automata, we established an undetectable watermarking scheme.

REFERENCES

Scott Aaronson. My ai safety lecture for ut effective altruism, November 2022. URL https://scottaaronson.blog/?p=6823. Accessed May 2023.

Dana Angluin and Michael Kharitonov. When won't membership queries help? In *Proceedings of the twenty-third annual ACM symposium on Theory of computing*, pp. 444–454, 1991.

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

Mikhail J Atallah, Victor Raskin, Christian F Hempelmann, Mercan Karahan, Radu Sion, Umut Topkara, and Katrina E Triezenberg. Natural language watermarking and tamperproofing. In *International workshop on information hiding*, pp. 196–212. Springer, 2002.

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

Miranda Christ and Sam Gunn. Pseudorandom error-correcting codes. In *Annual International Cryptology Conference*, pp. 325–347. Springer, 2024.

Miranda Christ, Sam Gunn, and Or Zamir. Undetectable watermarks for language models. In *The Thirty Seventh Annual Conference on Learning Theory*, pp. 1125–1139. PMLR, 2024.

Alexander Clark and Franck Thollard. Pac-learnability of probabilistic deterministic finite state automata. *Journal of Machine Learning Research*, 5(May):473–497, 2004.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Noah Golowich and Ankur Moitra. Edit distance robust watermarks via indexing pseudorandom codes. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=FZ45kf5pIA.

Zhengmian Hu, Lichang Chen, Xidong Wu, Yihan Wu, Hongyang Zhang, and Heng Huang. Unbiased watermark for large language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=uWVC5FVidc.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023.

Michael Kearns, Yishay Mansour, Dana Ron, Ronitt Rubinfeld, Robert E Schapire, and Linda Sellie. On the learnability of discrete distributions. In *Proceedings of the twenty-sixth annual ACM symposium on Theory of computing*, pp. 273–282, 1994.

John Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, and Tom Goldstein. A watermark for large language models. In *International Conference on Machine Learning*, pp. 17061–17084. PMLR, 2023.

Rohith Kuditipudi, John Thickstun, Tatsunori Hashimoto, and Percy Liang. Robust distortion-free watermarks for language models. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=FpaCL1MO2C.

Mathieu Mercadier and Frank Strobel. A one-sided vysochanskii-petunin inequality with financial applications. *European Journal of Operational Research*, 295(1):374–377, 2021.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020.

Dana Ron, Yoram Singer, and Naftali Tishby. On the learnability and usage of acyclic probabilistic finite automata. In *Proceedings of the eighth annual conference on Computational learning theory*, pp. 31–40, 1995.

Irene Solaiman, Miles Brundage, Jack Clark, Amanda Askell, Ariel Herbert-Voss, Jeff Wu, Alec Radford, Gretchen Krueger, Jong Wook Kim, Sarah Kreps, et al. Release strategies and the social impacts of language models. *arXiv preprint arXiv:1908.09203*, 2019.

Sebastiaan A Terwijn. On the learnability of hidden markov models. In *International Colloquium on Grammatical Inference*, pp. 261–268. Springer, 2002.

Mercan Topkara, Cuneyt M Taskiran, and Edward J Delp III. Natural language watermarking. In *Security, Steganography, and Watermarking of Multimedia Contents VII*, volume 5681, pp. 441–452. SPIE, 2005.

Robert A Wagner. Order-n correction for regular languages. *Communications of the ACM*, 17(5): 265–268, 1974.

Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. Defending against neural fake news. *Advances in neural information processing systems*, 32, 2019.

Xuandong Zhao, Prabhanjan Vijendra Ananth, Lei Li, and Yu-Xiang Wang. Provable robust watermarking for AI-generated text. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=SsmT8aO45L.

A LIMITATIONS

Our work has several limitations that suggest directions for future research. First, the lack of tight analytical bounds on the p-value remains a challenge, as the edit distance metric is inherently difficult to analyze. Second, the detection algorithm requires knowledge of the private key, which limits detection to only the key holder. Developing an asymmetric decoding and detection system could improve security.

B ADDITIONAL EMPIRICAL RESULTS

B.1 VARYING TEXT LENGTHS WITH SUPPLEMENTAL EVALUATION METRICS

Tables 2 and 3 present ROC-AUC results across varying text lengths, while Tables 4 and 5 report the corresponding TPR@1%FPR results. Overall, WEPA ($d = 1$) achieves performance comparable to Seq on both metrics, with WEPA ($d = 2$) performing even better. Although the Unbiased method attains the highest scores on these metrics, it is not robust to edit-based attacks.

Text length	**G/R**	**Seq**	**Unbiased**	**WEPA** ($d = 1$)	**WEPA** ($d = 2$)
5	0.829	0.751	0.790	0.765	0.764
10	0.910	0.880	0.939	0.862	0.906
15	0.952	0.938	0.979	0.922	0.950
20	0.971	0.964	0.992	0.951	0.974
25	0.982	0.969	0.995	0.964	0.988
30	0.986	0.982	0.998	0.980	0.990
35	0.988	0.989	0.998	0.985	0.992
40	0.991	0.991	0.999	0.991	0.993
45	0.992	0.993	0.999	0.992	0.996
50	0.993	0.995	0.999	0.996	0.998

Table 2: ROC-AUC results varying text lengths on LLaMA-3B.

Text length	**G/R**	**Seq**	**Unbiased**	**WEPA** ($d = 1$)	**WEPA** ($d = 2$)
5	0.812	0.678	0.784	0.681	0.705
10	0.898	0.817	0.925	0.779	0.848
15	0.938	0.894	0.971	0.859	0.913
20	0.958	0.924	0.991	0.901	0.940
25	0.967	0.934	0.998	0.921	0.971
30	0.975	0.952	0.998	0.952	0.982
35	0.981	0.967	1.000	0.960	0.982
40	0.984	0.979	1.000	0.968	0.989
45	0.987	0.975	1.000	0.976	0.993
50	0.988	0.988	1.000	0.981	0.993

Table 3: ROC-AUC results varying text lengths on Mistral-7B.

B.2 VARYING TEXT LENGTHS ON LONG TEXTS

We evaluated WEPA and Seq on longer text sequences, as presented in Figure 4. For cases where the empirical p-values became exponentially small, direct computation was infeasible; instead, we report upper bounds estimated via Equation 15. Due to the looseness of this bound, these values are not directly comparable to G/R. Nonetheless, the results reveal consistent trends with those in Figure 2. Notably, WEPA ($d = 1$) achieves the strongest performance, while WEPA ($d = 2$) exhibits slightly lower accuracy compared to Seq.

B.3 PERPLEXITY EVALUATION

We evaluated the perplexity of each watermarking scheme at a text length of 512, and report median perplexity with a 90% confidence interval. The results are shown in Figure 5. Among the methods,

Text length	G/R	Seq	Unbiased	WEPA ($d = 1$)	WEPA ($d = 2$)
5	0.121	0.032	0.085	0.037	0.040
10	0.258	0.071	0.269	0.094	0.114
15	0.347	0.190	0.314	0.157	0.158
20	0.487	0.344	0.319	0.293	0.341
25	0.621	0.318	0.965	0.259	0.647
30	0.709	0.525	0.998	0.499	0.592
35	0.709	0.608	1.000	0.517	0.761
40	0.649	0.596	1.000	0.790	0.750
45	0.693	0.740	1.000	0.671	0.990
50	0.701	0.837	1.000	0.892	0.996

Table 4: TPR@1%FPR results varying text lengths on LLaMA-3B.

Text length	G/R	Seq	Unbiased	WEPA ($d = 1$)	WEPA ($d = 2$)
5	0.052	0.019	0.085	0.032	0.035
10	0.269	0.032	0.277	0.036	0.061
15	0.304	0.150	0.319	0.063	0.144
20	0.320	0.163	0.320	0.135	0.124
25	0.494	0.149	0.989	0.140	0.327
30	0.549	0.243	0.995	0.336	0.534
35	0.666	0.308	1.000	0.258	0.521
40	0.717	0.483	1.000	0.431	0.714
45	0.757	0.355	1.000	0.454	0.837
50	0.761	0.671	1.000	0.510	0.728

Table 5: TPR@1%FPR results varying text lengths on Mistral-7B.

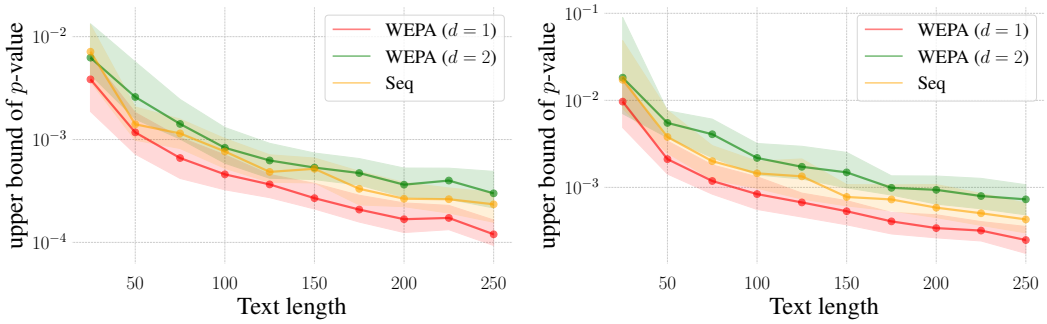

Figure 4: Median p-values varying text lengths on LLaMA-3B (left) and Mistral-7B (right).

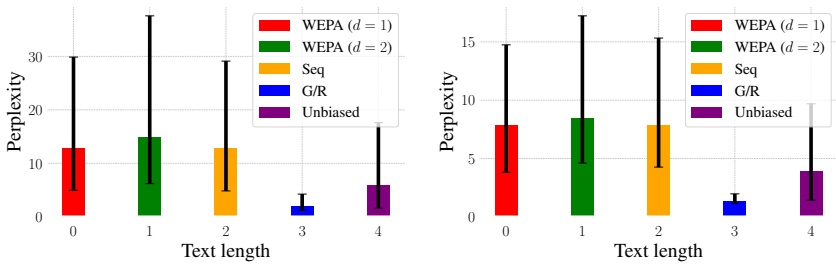

Figure 5: Perplexity of each watermarking scheme on LLaMA-3B (left) and Mistral-7B (right).

WEPA ($d = 2$) distorts the text the least. WEPA ($d = 1$) and Seq yield the same perplexity since they have the same generation process, but Seq heavily distorts the text.

B.4 Varying key length

We varied the key length (λ) from 2^4 to 2^{12} while keeping the text length fixed at 8, as shown in Figure 6. Consistent with the findings of Kuditipudi et al. (2024), we observe that the p-value increases with λ.

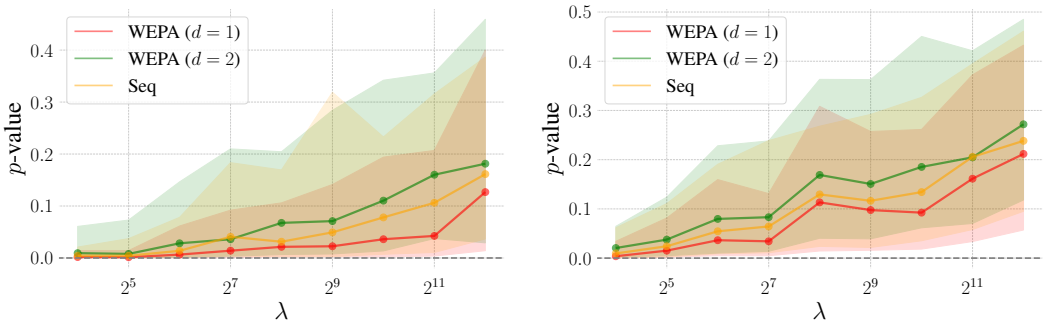

Figure 6: Median p-values varying security parameters on LLaMA-3B (left) and Mistral-7B (right).

B.5 Varying bitwidth

We varied the bitwidth (b) of WEPA from 2 to 12 while keeping the text length fixed at 8, as shown in Figure 7. The results indicate that a higher bitwidth improves performance with sacrifice of text diversity. The performance stabilizes when $b \geq 6$.

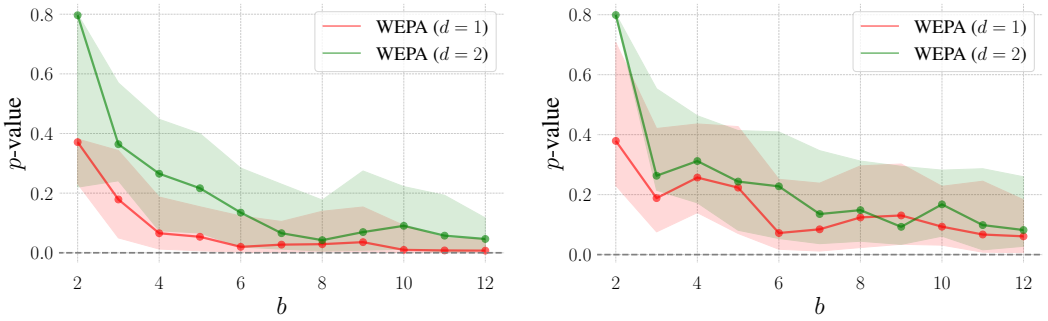

Figure 7: Median p-values varying number of bits on LLaMA-3B (left) and Mistral-7B (right).

B.6 Robustness to edit-based attack on Mistral-7B

We evaluated robustness against edit-based attacks on Mistral-7B, following the setup in Section 6.2, with results shown in Figure 8. Under a lower-entropy model, G/R demonstrates greater stability. Overall, these schemes exhibit performance similar to the LLaMA-3B model.

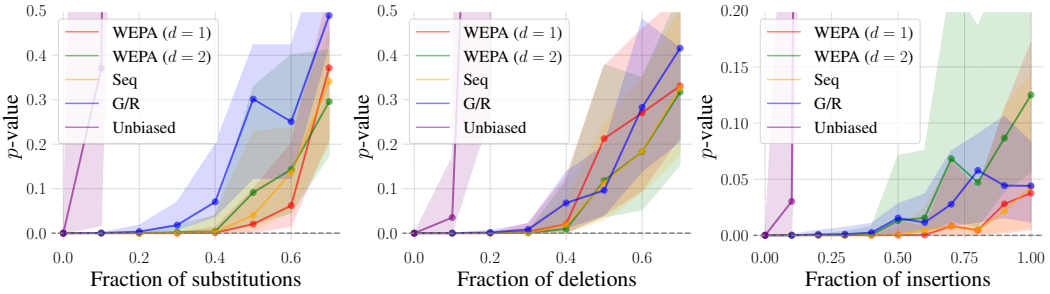

Figure 8: Median p-values under random substitution (left), deletion (middle), and insertion (right) attacks.

B.7 Numerical results of Section 6.1

Tables 6 and 7 list the numerical median p-values corresponding to Figure 2.

Text length	**G/R**	**Seq**	**Unbiased**	**WEPA** ($d = 1$)	**WEPA** ($d = 2$)
4	0.438	0.233	9.553	0.286	0.212
8	0.169	0.062	0.412	0.153	0.018
12	0.020	0.003	0.013	0.011	0.001
16	0.010	0.001	0.000	0.003	0.000
20	0.005	0.000	0.000	0.000	0.000

Table 6: Median p-values varying text lengths on LLaMA-3B.

Text length	**G/R**	**Seq**	**Unbiased**	**WEPA** ($d = 1$)	**WEPA** ($d = 2$)
4	0.438	0.235	8.145	0.285	0.295
8	0.169	0.100	0.692	0.136	0.106
12	0.078	0.011	0.034	0.061	0.007
16	0.010	0.015	0.002	0.029	0.002
20	0.005	0.001	0.000	0.003	0.000

Table 7: Median p-values varying text lengths on Mistral-7B.

C Experiment Setup

C.1 Computation resources

Experiments were run on an internal cluster with dual AMD EPYC 7453 28-core CPUs (56 cores total), 1 TiB RAM, and 8 NVIDIA GeForce GPUs (24 GiB each, CUDA 11.6). Peak GPU memory usage reached approximately 15 GiB per device. The system provided ample memory and storage, with over 800 GiB RAM available during runs. However, our experiments primarily rely on GPU resources for language model inference, and do not require high-performance CPUs.

C.2 Hyperparameter settings

We detail the hyperparameter configurations used for all methods evaluated in our experiments. For WEPA, we set the key length λ to 256. For the cyclic key sequence watermarking baseline (Seq), we used a key sequence length of 256 and applied exponential minimum sampling with a soft Levenshtein cost parameter $\gamma = 0$. For the green-red set watermarking baseline (G/R) Kirchenbauer

et al. (2023), we the green set fraction is set at 0.25 and the logit bias $\delta = 2$. For the unbiased watermarking method of Hu et al. (2024), we used δ-reweighting.

For WEPA, unless specified, we used the key length of λ 256 and `float32` type to mimic the behavior when b is sufficiently large. For Levenshtein distance, we set the deletion cost $\gamma_d = 0$ and insertion cost $\gamma_i = 2$. Notably, γ_i cannot be too small as a lower insertion cost allows tokens to align with arbitrary states with minimal penalty. For each experiment, we computed the z-scores for watermarked text across 100 samples.

D Implementation details

D.1 Implementation of Levenshtein distance calculation

Algorithm 2 presents the dynamic programming for computing the Levenshtein distance used in WEPA. The $\text{cost}(\cdot, \cdot)$ array stores the precomputed costs between substates and tokens defined by Equation 11.

Algorithm 2 Levenshtein Distance Calculation with Dynamic Programming

procedure LevenshteinDistance($\boldsymbol{y}$, λ, d, $\text{cost}(\cdot, \cdot)$, γ_d, γ_i)
  $m \leftarrow |\boldsymbol{y}|$
  $f, g \leftarrow$ array of size λ initialized to 0 ▷ Assume f and g are cyclic.
  for $i \leftarrow 1..m$ **do**
    for $u \leftarrow 0..\lambda - 1$ **do**
      $d_0 \leftarrow \text{cost}(u, y_{i-1})$
      $f_u \leftarrow g_u + \gamma_d$ ▷ Deletion
      for $v \leftarrow u - d..u - 1$ **do**
        $f_u \leftarrow \min(f_u, g_v + d_0)$ ▷ Substitution
    $u^* \leftarrow \arg\min(f_:)$
    for $u \leftarrow u^* + 1..\lambda - 1$ **do**
      for $v \leftarrow u - d..u - 1$ **do**
        $f_u \leftarrow \min(f_u, f_v + \gamma_i)$ ▷ Insertion
    for $u \leftarrow 0..u^* - 1$ **do**
      for $v \leftarrow u - d..u - 1$ **do**
        $f_u \leftarrow \min(f_u, f_v + \gamma_i)$ ▷ Insertion
    $f, g \leftarrow g, f$
  return $\min(f_:)$

While the innermost loop can be optimized using a monotonic queue for a complexity of $\mathcal{O}(m\lambda)$, we do not carry out this optimization in our implementation in practice since the degree d is usually small.

E Additional analysis of existing watermarking schemes

In this section, we provide the alternative analysis of existing watermarking schemes through the lens of probabilistic automata (PA).

E.1 Watermarking schemes under uniform models

We assume a model that follows a uniform distribution over $\mathcal{V}$ so that $\mathsf{Model}(\cdot) = \text{Cat}(1/|\mathcal{V}|, \ldots, 1/|\mathcal{V}|)$, where Cat denotes the categorical distribution. Then we proceed to analyze the output distribution of existing watermarking schemes under uniform models.

k-gram-based watermarking. k-gram-based watermarking has been introduced in many works (Kirchenbauer et al., 2023; Zhao et al., 2024; Aaronson, 2022). Generally speaking, the core concept is to condition the next token on a window of k prior tokens. A hash function $h : \mathcal{K} \times \mathcal{V}^k \to \mathbb{Z}$ generates the noise based on the context of the k prior tokens. Then Φ computes the hash and alters the distribution and generates a variable:

$$\xi_i \sim \Phi(\boldsymbol{y}_{i-k:i-1}) = \Phi'(h_{\mathsf{sk}}(\boldsymbol{y}_{i-k:i-1})), \tag{22}$$

where $\Phi' : \mathbb{Z} \to \Delta(\Xi)$. The noise, ξ_i may be modeled as a uniform variable from $[0, 1]$ used for inverse transform sampling with a partition over the vocabulary (Kirchenbauer et al., 2023; Zhao et al., 2024), or a Gumbel variable (Aaronson, 2022).

Under uniform models, the output of a watermarking scheme can be represented by a PA $\mathcal{M} = (Q, \Sigma, \delta, \pi_0, \pi_f)$, with the state space $Q = \mathcal{V}^k$ and $\Xi = \mathcal{V}$, the transition function $\delta : Q \times \Sigma \times Q \to [0, 1]$ is defined as

$$\begin{aligned} \delta(\boldsymbol{y}_{i-k:i-1}, y_i, \boldsymbol{y}_{i-k+1:i}) = \mathbb{P}(Y = y_i \mid \boldsymbol{y}_{i-k:i-1}), \\ Y \sim \Phi(\boldsymbol{y}_{i-k:i-1}). \end{aligned} \tag{23}$$

As each $\boldsymbol{y}_{i-k:i-1}$ and y_i uniquely determines $\boldsymbol{y}_{i-k+1:i}$, $\mathcal{M}$ is a PDFA.

Cyclic key sequence watermarking. Kuditipudi et al. (2024) introduced a watermarking scheme that uses a cyclic sequence of noise $\ldots \xi_{\lambda-1}\xi_0\xi_1 \ldots \xi_{\lambda-1}\xi_0 \ldots$ starting from random position in watermarks within generated text, where each $\xi_i \in [0, 1]$ for inverse transform sampling or $\xi_i \in [0, 1]^{|\mathcal{V}|}$ for exponential minimum sampling. The noise sequence is referred to as the key sequence.

Since the key sequence is cyclic, the output under uniform models also follows a repeated structure and can be recognized by a PDFA with at most 4λ states, which can be constructed with simliar ideas as suffix automata. We defer the construction in Appendix H.2.

E.2 Detectability of watermarking schemes

Definition 11 (KL-PAC Learnability)**.** Given a class of stochastic languages or distributions $\mathcal{C}$ over Σ^*, an algorithm $\mathcal{A}$ *KL-Probably Approximately Correctly (KL-PAC)-learns* $\mathcal{C}$ if there exists a polynomial q such that for all $c \in \mathcal{C}$, all $\epsilon > 0$ and $\delta > 0$, $\mathcal{A}$ is given a sample S_m and produces a hypothesis H satisfying

$$\mathbb{P}\left[D_{\mathsf{KL}}(c||H) > \epsilon\right] < \delta \tag{24}$$

whenever $m > q(1/\epsilon, 1/\delta, |c|)$, where $|c|$ is some measure of the complexity of the target. The algorithm runs in time polynomial in m plus the total length of the strings in S_m.

Theorem 3. *For any $\mu > 0$, any watermarking scheme the output of which under a uniform model represented by a μ-distinguishable PDFA with polynomial many states is detectable.*

Theorem 3 provides an alternative perspective to show that the watermarks proposed by Kirchenbauer et al. (2023); Zhao et al. (2024); Aaronson (2022); Kuditipudi et al. (2024) are all detectable and even spoofable.

F Proof of technical results

F.1 Proof of Theorem 1

Proof. We begin we noticing that

$$\mathrm{Adv}_{\mathcal{W}}^{\mathsf{IND\text{-}OT}}(\mathcal{A}) = \left|\Pr\left[b = b'\right] - \frac{1}{2}\right| = 0 \quad \Leftrightarrow \quad \Pr\left[b = b'\right] = \frac{1}{2}. \tag{25}$$

Necessity. Fix an arbitrary output $\boldsymbol{y} \in \mathcal{L}$ for a given key $\mathsf{sk} \in K$. Suppose an adversary $\mathcal{A}$ attempts to distinguish between the original and watermarked models. The adversary may partition the output space $\mathcal{L}$ into two disjoint sets:

$$\mathcal{L}_0, \quad \mathcal{L}_1, \tag{26}$$

where $\mathcal{L}_0 \cap \mathcal{L}_1 = \varnothing$ and $\mathcal{L}_0 \cup \mathcal{L}_1 = \mathcal{L}$. The adversary's strategy is to output $b' = 0$ if $\boldsymbol{y} \in \mathcal{L}_0$ and $b' = 1$ otherwise. The probability of correctly guessing b is then given by:

$$\begin{aligned} \Pr[b = b'] &= \Pr[b' = 0 \mid b = 0] \Pr[b = 0] + \Pr[b' = 1 \mid b = 1] \Pr[b = 1] \\ &= \frac{1}{2}\left(\Pr[\boldsymbol{y} \in \mathcal{L}_0 \mid K = \mathsf{sk}, b = 0] + \Pr[\boldsymbol{y} \in \mathcal{L}_1 \mid K = \mathsf{sk}, b = 1]\right). \end{aligned} \tag{27}$$

Since the watermarking scheme $\mathcal{W}$ is assumed to be perfectly distortion-free, we have:

$$\Pr[Y = \boldsymbol{y}] = \mathbb{E}_{\mathsf{sk} \leftarrow \mathsf{Gen}(1^\lambda)}\Big[\Pr[Y = \boldsymbol{y} \mid K = \mathsf{sk}]\Big]. \tag{28}$$

Taking expectations over the key sk, it follows that:

$$\Pr[\boldsymbol{y} \in \mathcal{L}_0] = \mathbb{E}_{\mathsf{sk} \leftarrow \mathsf{Gen}(1^\lambda)} \left[\Pr[\boldsymbol{y} \in \mathcal{L}_0 \mid K = \mathsf{sk}] \right], \tag{29}$$

with a similar expression for $\mathcal{L}_1$. By the law of total probability, we obtain:

$$\Pr[\boldsymbol{y} \in \mathcal{L}_0] + \Pr[\boldsymbol{y} \in \mathcal{L}_1] = 1. \tag{30}$$

Substituting into the expression for $\Pr[b = b']$ and noting that b is independent of sk, we conclude:

$$\Pr[b = b'] = \frac{1}{2} \left(\Pr[\boldsymbol{y} \in \mathcal{L}_0] + \Pr[\boldsymbol{y} \in \mathcal{L}_1] \right) = \frac{1}{2}. \tag{31}$$

Since the adversary cannot achieve a probability of success greater than $\frac{1}{2}$, we conclude:

$$\mathrm{Adv}_{\mathcal{W}}^{\mathsf{IND\text{-}OT}}(\mathcal{A}) = \left| \Pr[b = b'] - \frac{1}{2} \right| = 0. \tag{32}$$

Thus, the adversary gains no distinguishing advantage, proving that the watermarking scheme $\mathcal{W}$ is perfectly indistinguishable.

Sufficiency. Conversely, assume that the watermarking scheme $\mathcal{W}$ is *not* perfectly distortion-free. Then there exists a language model with output distribution $\mathcal{L}$ and some $\boldsymbol{y} \in \mathcal{L}$ such that:

$$\Pr[Y = \boldsymbol{y}] \;\neq\; \mathbb{E}_{\mathsf{sk} \leftarrow \mathsf{Gen}(1^\lambda)} \big[\Pr[Y = \boldsymbol{y} \mid K = \mathsf{sk}] \big]. \tag{33}$$

Define the disjoint sets:

$$\mathcal{L}_0 \;=\; \{ \boldsymbol{y} \in \mathcal{L} \mid \Pr[Y = \boldsymbol{y}] \;>\; \mathbb{E}_{\mathsf{sk} \leftarrow \mathsf{Gen}(1^\lambda)} \big[\Pr[Y = \boldsymbol{y} \mid K = \mathsf{sk}] \big] \}, \quad \mathcal{L}_1 \;=\; \mathcal{L} \setminus \mathcal{L}_0. \tag{34}$$

Consider an adversary $\mathcal{A}$ that outputs $b' = 0$ when $\boldsymbol{y} \in \mathcal{L}_0$ and $b' = 1$ when $\boldsymbol{y} \in \mathcal{L}_1$. The probability of a correct guess is given by:

$$\Pr[b = b'] = \frac{1}{2} \left(\Pr[\boldsymbol{y} \in \mathcal{L}_0 \mid b = 0] + \Pr[\boldsymbol{y} \in \mathcal{L}_1 \mid b = 1] \right). \tag{35}$$

Since $\boldsymbol{y} \in \mathcal{L}_0$ appears more frequently in the original model's distribution than in the watermarked model's, and vice versa for $\mathcal{L}_1$, it follows that:

$$\Pr[b = b'] > \frac{1}{2}. \tag{36}$$

Thus, the adversary gains a nonzero advantage, contradicting the assumption that $\mathcal{W}$ is perfectly distortion-free, which completes the proof. □

F.2 Proof of Theorem 3

Proof. Since any μ-distinguishable PDFA with a polynomial number of states and a bound on the expected length of strings generated from any state is KL-PAC-learnable (Clark & Thollard, 2004). That is, given sufficiently many samples from the distribution induced by a μ-distinguishable PDFA, an efficient learning algorithm can approximate the distribution arbitrarily well.

Suppose, for the sake of contradiction, that there exists a watermarking scheme $\mathcal{W}$ that produces outputs indistinguishable from those of an unwatermarked uniform model represented by a μ-distinguishable PDFA $\mathcal{M}$. That is, for any adversary $\mathcal{A}$,

$$\mathrm{Adv}_{\mathcal{W}}^{\mathrm{ind}}(\mathcal{A}) = \left| \Pr[b = b'] - \frac{1}{2} \right| = 0. \tag{37}$$

This implies that no efficient adversary can distinguish between the original model $\mathcal{M}$ and the watermarked model $\mathcal{M}'$ with non-trivial probability.

However, since $\mathcal{M}$ is μ-distinguishable, it is KL-PAC-learnable. Thus, there exists a polynomial-time algorithm $\mathcal{L}$ that, given polynomially many samples from the output distribution of either $\mathcal{M}$ or $\mathcal{M}'$, can learn a hypothesis h that approximates the true distribution up to any desired accuracy ϵ.

Consider the following adversary $\mathcal{A}$ that attempts to distinguish between $\mathcal{M}$ and $\mathcal{M}'$:

1. Draw $m = \text{poly}(1/\epsilon, 1/\delta)$ samples from the given black-box model.
2. Run the KL-PAC-learning algorithm $\mathcal{L}$ to obtain a hypothesis h that approximates the underlying distribution.
3. Compute the likelihood of the observed samples under both the learned approximation of $\mathcal{M}$ and $\mathcal{M}'$.
4. Output $b' = 0$ if the samples are closer to the expected distribution of $\mathcal{M}$ and $b' = 1$ otherwise.

Since $\mathcal{M}$ is μ-distinguishable, every pair of states in $\mathcal{M}$ induces suffix distributions that differ by at least μ in ℓ_∞ norm, and $\mathcal{M}$ terminates with $m = \text{poly}(\lambda)$ steps. If the watermarking scheme $\mathcal{W}$ alters the output distribution, then by definition of μ-distinguishability, it must introduce a statistically significant difference that is detectable given polynomially many samples.

Therefore, the KL-PAC-learning algorithm $\mathcal{L}$ enables the adversary $\mathcal{A}$ to distinguish between the original and watermarked models with probability strictly greater than $\frac{1}{2} + \gamma$ for some $\gamma > 0$. This contradicts our initial assumption that the watermarking scheme is perfectly indistinguishable.

Hence, we conclude that for any $\mu > 0$, any watermarking scheme applied to a uniform model represented by a μ-distinguishable PDFA with a polynomial number of states is necessarily detectable. □

Since all watermarking schemes proposed in Kirchenbauer et al. (2023); Zhao et al. (2024); Aaronson (2022); Kuditipudi et al. (2024) can be represented as a PDFA, and each pair of states in these automata is distinguishable (except in the unlikely case where two randomly chosen partitions are identical or the corresponding uniform random variables are nearly indistinguishable) it follows that all such watermarking schemes are necessarily detectable.

G UNDETECTABLE WATERMARK CONSTRUCTION

Before presenting the watermark construction, we introduce the necessary definitions and assumptions.

Definition 12 (PAC Learnability)**.** Let $\mathcal{C}$ be a concept class consisting of Boolean functions $c : \mathcal{X} \to \{0, 1\}$ over some instance space $\mathcal{X}$. We say that an algorithm $\mathcal{A}$ *Probably Approximately Correctly (PAC)-learns* $\mathcal{C}$ if there exists a polynomial q such that for all target concepts $c \in \mathcal{C}$, all distributions $\mathcal{D}$ over $\mathcal{X}$, and all $\epsilon, \delta > 0$, the algorithm $\mathcal{A}$, given access to i.i.d. examples drawn from $\mathcal{D}$ labeled by c, outputs a hypothesis h such that

$$\Pr\left[\ \Pr_{x\sim\mathcal{D}}\left[h(x) \neq c(x)\right] > \epsilon\ \right] < \delta \tag{38}$$

whenever the number of examples $m > q(1/\epsilon, 1/\delta, |c|)$, where $|c|$ denotes a measure of the complexity of the target concept. The algorithm runs in time polynomial in m plus the size of the input examples.

Assumption 4 (Sparsely Learning Parities with Noise)**.** Let $\mathcal{F} = (F_n)_{n\in\mathbb{N}}$ be the class of parity functions $F_n = \{f_s(x) = \bigoplus_{i\in[n]} x_i s_i \mid s \in \{0,1\}^n,\ \|s\|_1 = \log n\}$, where $x \in \{0,1\}^n$. Each f_s computes the parity of a logarithmic-sized subset of the input, defined by the support of s. No polynomial-time algorithm can PAC learn $\mathcal{F}$ under classification noise rate $q = 1/3$.

Definition 13 (Entropy Bound of Language Model)**.** Let $(y_1, y_2, \ldots, y_n)$ be a sequence of tokens generated by a language model over a finite vocabulary $\mathcal{V}$. The model assigns a conditional distribution $p(y_i \mid \boldsymbol{y}_{1:i-1})$ at each position i. We define the entropy bound of a language model as

$$\underline{H}(\mathsf{Model}) = \inf\left\{\mathbb{E}_{i\sim\mathcal{U}[n]}\left[H\left(y_i \mid \boldsymbol{y}_{1:i-1}\right)\right]\right\}, \tag{39}$$

where $H(y_i \mid \boldsymbol{y}_{1:i-1})$ denotes the conditional entropy and the expectation is taken over a uniformly random index $i \in [n]$.

Definition 14 (Weak Pseudorandom Function Family)**.** A family of functions $\mathcal{F} = \{f_s : \{0,1\}^n \to \{0,1\}\}_{s\in\{0,1\}^n}$ is a *weak pseudorandom function (PRF) family* if for every probabilistic polynomial-time adversary $\mathcal{A}$, the distinguishing advantage

$$\left| \Pr_{s\sim\{0,1\}^n}[\mathcal{A}^{f_s}(1^n) = 1] - \Pr_{f\sim\mathcal{U}}[\mathcal{A}^{f}(1^n) = 1] \right| \leq \text{negl}(n), \tag{40}$$

where $\mathcal{U}$ denotes the uniform distribution over all functions $f : \{0,1\}^n \to \{0,1\}$, and the oracle queries to $\mathcal{A}$ are drawn uniformly at random from $\{0,1\}^n$.

Proposition 5 (From PAC Unlearnability to Weak Pseudorandomness)**.** *Let $\mathcal{F} = (\mathcal{F}_n)_{n\in\mathbb{N}}$ be a class of functions that is not PAC learnable under classification noise rate q. Then, for $n(\lambda) = \lambda$, the sequence of functions $(\mathcal{F}_{n(\lambda)})_\lambda$ constitutes a weak pseudorandom function family with noise level q.*

G.1 Full Statement and Proof of Theorem 2

Reduction to a binary vocabulary. For analytical convenience, we reduce the vocabulary of $\mathcal{V}$ to a binary vocabulary $\{0,1\}$. Any categorical distribution over $\mathcal{V}$ with $|\mathcal{V}|$ can be equivalently represented by a binary distribution over $\{0,1\}^{\lceil \log_2 |\mathcal{V}| \rceil}$ via entropy-preserving encoding. That is, given logits $p \in \Delta^{|\mathcal{V}|-1}$, we define a mapping to binary sequences where each symbol $v \in \mathcal{V}$ is assigned a unique bitstring $b(v) \in \{0,1\}^{\lceil \log_2 |\mathcal{V}| \rceil}$, and generation proceeds bit by bit using the induced marginal distributions. This reduction preserves perplexity and generation quality up to negligible statistical error. We therefore assume, without loss of generality, that the model outputs binary tokens.

We present the full statement of Theorem 2 as follows.

Theorem 6. *There exists an undetectable watermarking scheme with generation length $n = \Omega(\lambda^3 \log \lambda)$, entropy bounded by $\underline{H}(\mathsf{Model}) > 2 + \log_2(1-q) - \frac{3-4q}{4(1-q)} \log_2(3-4q)$ that can be represented by a PA, if sparse LPN is hard with noise level of $0 < q < \frac{1}{2}$.*

Proof. Let $f_{\boldsymbol{s}}(\boldsymbol{x}) = \boldsymbol{s} \cdot \boldsymbol{x} \bmod 2$ be a parity function with noise level of q on the support of a secret key $\boldsymbol{s} \in \{0,1\}^\lambda$, and suppose a language model with binary output alphabet $\{0,1\}$.

At each token position i, we generate a pair of Gumbel noise variables (μ_0, μ_1) sampled from $U[0,1]^2$, whose binary representations are given by Equation 9.

If $(\lambda+1) \nmid i$, the pair (μ_0, μ_1) is left unmodified. Otherwise, when $(\lambda+1) \mid i$, we extract a watermark bit x_i via the indicator

$$x_i = \mathbf{1}\left[\mu_0 < \mu_1\right], \tag{41}$$

compute the parity bit $b = f_{\boldsymbol{s}}(\boldsymbol{x})$ over the current watermark buffer $\boldsymbol{x}$, and enforce $x_i = b$ by outputting (μ_0, μ_1) with probability $1-q$ if $b = 0$ or with probability q and $b = 1$, otherwise outputting (μ_1, μ_0). This is easily accomplished by a PA by counting the number of bits of $\boldsymbol{s}$. The decoder of Equation 8 is used for each step to generating y_i.

The detector identifies all positions i such that $(\lambda+1) \mid i$ and reconstructs the corresponding Gumbel pairs (μ_0, μ_1) used during generation. It computes the bit

$$x_i = \mathbf{1}[\mu_0 < \mu_1], \tag{42}$$

and compares it against the output token $y_i \in \{0,1\}$. For each such position, let $z_i = \mathbf{1}[x_i = y_i]$ denote a match indicator. Over $t = \Omega(n/\lambda)$ such comparisons, the detector computes the empirical match rate:

$$\hat{p} = \frac{1}{t}\sum_{i=1}^{t} z_i. \tag{43}$$

The detector accepts (i.e., concludes the presence of a watermark) if $\hat{p} \geq \frac{1}{2} + \theta$ for some threshold $\theta = \Omega(\lambda^{-1})$, and rejects otherwise.

Undetectability. Let D_S denote the distribution over input-output pairs induced by our construction, where the input $\boldsymbol{x} \in \{0,1\}^\lambda$ is uniformly random and the output bit is given by $f_{\boldsymbol{s}}(\boldsymbol{x})$ with noise

rate q. This corresponds to the standard *noisy parity* distribution: with probability $1-q$, the output is $f_{\boldsymbol{s}}(\boldsymbol{x})$, and with probability q, it is flipped.

Suppose there exists an algorithm $\mathcal{A}$ that KL-PAC-learns the class of such distributions, i.e., for any $\epsilon > 0$ and $\delta > 0$, given m samples drawn from D_S, the algorithm produces a hypothesis distribution $\hat{D}$ such that

$$\Pr\left[D_{\mathsf{KL}}(D_S \,\|\, \hat{D}) > \epsilon\right] < \delta \tag{44}$$

whenever $m > q(1/\epsilon, 1/\delta, |D_S|)$ for some polynomial q.

Now consider a random input $\boldsymbol{x} \in \{0,1\}^\lambda$. Since the output bit is either $f_{\boldsymbol{s}}(\boldsymbol{x})$ or its complement, the two candidate outputs are $\boldsymbol{x}0$ and $\boldsymbol{x}1$. A hypothesis $\hat{D}$ that approximates D_S in KL-divergence necessarily assigns higher likelihood to the correct label in expectation. Therefore, by comparing the probabilities $\hat{D}(\boldsymbol{x}0)$ and $\hat{D}(\boldsymbol{x}1)$, one can predict $f_{\boldsymbol{s}}(\boldsymbol{x})$ with probability at least $1-\epsilon$.

This yields an efficient algorithm that predicts noisy parity with non-negligible advantage, contradicting the Noisy Parity Assumption.

Finally, by Proposition 5, the parity function $f_{\boldsymbol{s}}$ can be computed by a probabilistic automaton, and the corresponding output distribution D_S serves as a pseudorandom function. In particular, since D_S is not KL-PAC-learnable under the hardness of sparse LPN, it is indistinguishable from a uniform distribution by any efficient adversary.

Therefore, no efficient detector can distinguish watermarked text from unwatermarked text with non-negligible advantage.

Completeness. Let $p_0 \in (0, 1/2]$ denote the smallest possible bias such that the binary entropy of y_i satisfies

$$H(y_i) = -p_0 \log_2 p_0 - (1-p_0)\log_2(1-p_0). \tag{45}$$

Consider the probability that the sampled watermark bit x_i matches the LM output y_i when $i \mid (\lambda+1)$. Since $x_i = b$ with probability $1-q$, and is flipped with probability q (due to the noise in $f_{\boldsymbol{s}}$), we have the following bound:

$$\begin{aligned}
\Pr\left[x_i = y_i\right] &\geq \left[2p_0\left(\frac{1}{2}\right)^{\log\lambda}(1-q) + \frac{1}{2}\left(1 - \left(\frac{1}{2}\right)^{\log\lambda}\right)\right] \\
&= \frac{1}{2} + \lambda^{-1}\cdot\left(2p_0 - 2qp_0 - \frac{1}{2}\right) \\
&= \frac{1}{2} + c\lambda^{-1},
\end{aligned} \tag{46}$$

where $c := 2p_0 - 2qp_0 - \frac{1}{2} > 0$ by the entropy bound.

Then we have the following Chernoff bound:

$$\Pr\left[\hat{p} < \frac{1}{2} + \theta\right] \leq \exp\left(-2t(c\lambda^{-1} - \theta)^2\right) = \exp\left(-\frac{tc^2}{2\lambda^2}\right) = \mathrm{negl}(\lambda). \tag{47}$$

Soundness. For unwatermarked text, assuming the Gumbel samples and tokens are independent (i.e., no watermarking mechanism was used), the match indicators z_i are i.i.d. unbiased coin flips:

$$\Pr[x_i = y_i] = \frac{1}{2}. \tag{48}$$

Then by Hoeffding's inequality,

$$\Pr\left[\hat{p} \geq \tfrac{1}{2} + \theta\right] \leq \exp(-2t\theta^2) = \exp(-\Omega(\log\lambda)) = \mathrm{negl}(\lambda), \tag{49}$$

which indicates that the detector rejects unwatermarked text with overwhelming probability.

□

H ILLUSTRATIONS OF CONSTRUCTIONS

H.1 ILLUSTRATION OF A SUBORDINATE PROBABILISTIC AUTOMATON

To complement the formal description in Section 4.2, we provide an illustration of a subordinate probabilistic automaton (sub-PA) used in our watermarking framework. The sub-PA is responsible for generating the binary noise vectors $\mu_i \in \{0,1\}^c$ across vocabulary positions, as defined in the main text.

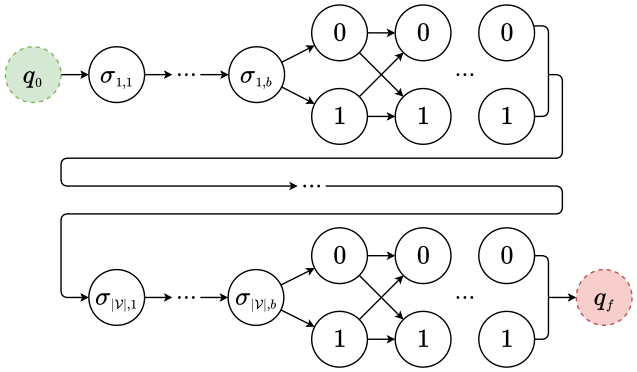

Figure 9: Illustration of a subordinate probabilistic automaton (sub-PA). Each layer encodes a binary vector μ_i, beginning from the initial state q_0 and terminating at q_f. Transitions within each layer encode bitwise states, branching into parallel Boolean paths $\iota_{i,j}$ and $\hat{\iota}_{i,j}$, which represent 0 and 1 respectively. Inter-layer transitions connect the terminal states of layer i to the initial states of layer $i+1$.

As shown in Figure 9, the sub-PA progresses from the initial state q_0 through $|\mathcal{V}|$ layers, each corresponding to a token in the vocabulary. The Boolean branching ensures exponential variability in possible noise sequences, while the overall number of automaton states remains polynomially bounded. This property is essential for both efficient sampling and tractable computation of edit distance during watermark detection.

H.2 AUTOMATON CONSTRUCTION FOR A CYCLIC STRING

We provide the construction of an automaton for a cyclic string in Algorithm 3. Given the input string s, the automaton has at most $4|s|$ states. Figure 10 illustrates an example PDFA that recognizes a substring of the repeated string of "abaa".

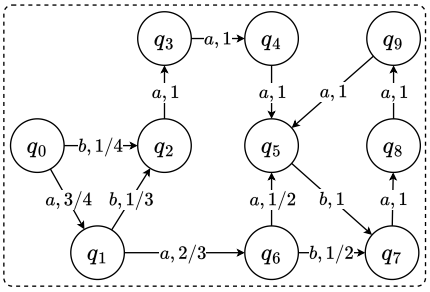

Figure 10: An example of a PDFA with probability that recognizes a substring of repeated string "abaa".

IMPACT STATEMENT

This study showcases recent advances in the field of Machine Learning, aimed at preventing the misuse of AI-generated content. Our technique endeavors to improve the security of text generation systems, thereby reducing the risk of misleading information and contributing to the establishment of ownership and copyright of AI-generated content. From an ethical perspective, this technology

Algorithm 3 Automaton Construction for a Cyclic String

Initialize: Let S be a set of states representing the suffix automaton, where each state S_i has:
- S_i.length: The length of the longest string ending at S_i.
- S_i.link: The suffix link, pointing to the longest proper suffix of the corresponding string that is also in the automaton.
- S_i.next: A transition function mapping each character c to the corresponding state.

Create the initial state S_0 with S_0.length $= 0$ and S_0.link $= -1$.
last $\leftarrow 0$ ▷ Tracks the last added state
function EXTEND(c)
    $p \leftarrow$ last
    Create a new state S_{cur} with S_{cur}.length $= S_p$.length $+ 1$
    while $p \neq -1$ and $c \notin S_p$.next **do**
        S_p.next$[c] \leftarrow$ cur
        $p \leftarrow S_p$.link
    if $p = -1$ **then**
        S_{cur}.link $\leftarrow 0$
    else
        $q \leftarrow S_p$.next$[c]$
        if S_p.length $+ 1 = S_q$.length **then**
            S_{cur}.link $\leftarrow q$
        else
            Create a new state S' with S'.length $= S_p$.length $+ 1$
            S'.next $\leftarrow S_q$.next
            S'.link $\leftarrow S_q$.link
            while $p \neq -1$ and S_p.next$[c] = q$ **do**
                S_p.next$[c] \leftarrow S'$
                $p \leftarrow S_p$.link
            S_q.link $\leftarrow S'$
            S_{cur}.link $\leftarrow S'$
    last $\leftarrow$ cur
function BUILDAUTOMATON(s)
    for each character c in s **do**
        EXTEND(c)
    idx $\leftarrow$ index of the newest state in the automaton
    for each character c in s **do**
        EXTEND(c)
    S_{last}.next$[s_0] \leftarrow$ idx

plays an important role in protecting the integrity of information dissemination on digital platforms and helps strengthen public trust in AI applications. While we recognize that these technologies may be used to restrict freedom of information or enforce censorship in some environments, we believe that their benefits in reducing misinformation outweigh the potential risks.