ON THE PAC LEARNABILITY OF DISTORTION-FREE LANGUAGE MODEL WATERMARKS

ABSTRACT

Distortion-free watermarking schemes for large language models embed watermarks into sampling randomness rather than token probabilities, achieving invisibility by preserving the exact output distribution of the base model. While such schemes are undetectable from a single sample, their security against learning-based adversaries has not been formally characterized. In this work, we provide a learning-theoretic analysis of distortion-free watermarking by studying the learnability of the induced watermark detector. We focus on cyclic distortion-free watermarking schemes with alignment-based detectors and prove that the associated detector class has low complexity and is PAC learnable from polynomially many labeled examples, such as detector queries. This result establishes that cyclic distortion-free watermarks are inherently vulnerable to learning-based attacks, despite their distributional invisibility and robustness to edit-distance perturbations. By leveraging probabilistic automata-based constructions introduced in Wang & Shang (2025) and standard cryptographic hardness assumptions, we demonstrate the existence of distortion-free watermarking schemes that are computationally hard to PAC learn.

1 INTRODUCTION

The widespread deployment of large language models (LLMs) has intensified the need for reliable mechanisms to identify machine-generated text. Watermarking has emerged as a promising approach for addressing concerns of provenance, misuse mitigation, and accountability in generative systems (Kirchenbauer et al., 2024; Aaronson & Kirchner, 2022). An effective watermark must satisfy several competing requirements: it should be statistically invisible to human users, robust to benign post-processing such as editing or paraphrasing, and secure against adversaries attempting to detect, remove, or spoof the watermark.

Early watermarking schemes for text generation modify the model's decoding distribution by introducing a secret bias into token selection. A prominent example is the k-gram–based approach of Kirchenbauer et al. (2024), which partitions the vocabulary into secret subsets and biases sampling toward preferred tokens. While such methods enable efficient detection, they necessarily distort the output distribution of the language model. This distortion makes them vulnerable to statistical detection and to learning-based attacks that exploit persistent distributional artifacts.

Recent work on distortion-free watermarking addresses this limitation by embedding the watermark into the sampling randomness rather than the model distribution itself (Kuditipudi et al., 2024). Using unbiased decoding procedures such as inverse transform sampling or exponential-minimum sampling, these schemes generate text that is exactly distributed according to the base language model, achieving information-theoretic invisibility against single-sample statistical tests. Detection is enabled by introducing secret correlations across tokens and employing alignment-based detectors that are robust to edit-distance perturbations, including insertions, deletions, substitutions, and cropping.

However, distortion-freeness alone does not guarantee security against stronger adversaries. While it rules out detection from a single sample, it does not prevent an adversary with access to multiple watermarked outputs or to the detector itself from learning sufficient structure to approximate the detector's decision rule. This observation raises a fundamental question that has remained largely unaddressed: to what extent are distortion-free watermarks secure against learning-based adversaries?

In this work, we study distortion-free watermarking through the lens of Probably Approximately Correct (PAC) learning. Rather than focusing on key recovery or direct distributional distinguishability, we analyze the learnability of the induced watermark detector. We focus on cyclic distortion-free watermarking schemes, which are widely used in practice due to their robustness to edit-distance perturbations and unknown alignment. We show that the alignment-based detectors associated with these schemes form a hypothesis class of low complexity and are PAC learnable from polynomially many labeled examples, such as detector queries. This provides a principled explanation for the vulnerability of cyclic distortion-free watermarks to learning-based attacks, despite their information-theoretic invisibility at the distributional level.

We further show that this vulnerability is not inherent to distortion-free watermarking itself. By leveraging recent automata-based formulations of watermarking (Wang & Shang, 2025) and classical hardness results from learning theory (Kearns et al., 1994), we demonstrate that more expressive constructions based on probabilistic nondeterministic finite automata can induce watermark distributions that are computationally hard to PAC learn under standard cryptographic assumptions. These results highlight learning complexity, rather than distributional distortion alone, as a central determinant of watermark security.

Our contributions are as follows:

1. We formalize security against learning-based adversaries for distortion-free watermarking by introducing PAC security of the detector as an additional desirable watermarking property.
2. We prove that the cyclic distortion-free watermarking scheme introduced in Kuditipudi et al. (2024)is PAC learnable from polynomially many labeled examples.
3. Building from Wang & Shang (2025), we demonstrate that distortion-free watermarking can achieve the aforementioned PAC learning-based security by constructing automata-based schemes whose induced distributions are computationally hard to PAC learn under standard cryptographic assumptions.

2 Related Work

Bias-based and hash-based watermarking. Early watermarking approaches for language models modify the decoding procedure by introducing a secret bias into token selection. A representative example is the k-gram–based watermarking scheme of Kirchenbauer et al. (2024), which uses a keyed hash of the recent context to partition the vocabulary into a “green list” and a “red list” and biases sampling toward green-list tokens. Detection proceeds by recomputing these partitions and applying a statistical hypothesis test to the observed token frequencies. Related ideas appear in earlier discussions of watermarking LLM outputs (Aaronson & Kirchner, 2022). While these methods are simple and computationally efficient, they necessarily distort the model’s output distribution, making them vulnerable to statistical detection and to learning-based attacks that exploit distributional artifacts.

Distortion-free watermarking. To avoid the inherent limitations of biased decoding, Kuditipudi et al. (2024) introduce the paradigm of distortion-free watermarking, in which the watermark is embedded into the sampling randomness rather than the model distribution. Their approach relies on unbiased decoding procedures such as inverse transform sampling and exponential-minimum sampling to preserve the exact output distribution of the base language model. Detection is achieved via alignment-based tests that measure statistical dependence between the generated text and a secret randomness sequence, enabling robustness to a wide range of edit-distance perturbations. While this framework resolves the problem of distributional distortion, it leaves open the question of security against adversaries with access to multiple samples or detector queries.

Automata-based formulations of watermarking. Recent work by Wang & Shang (2025) provides a unifying abstraction for watermarking schemes by modeling the generation of sampling randomness as a probabilistic automaton. In this framework, a watermark corresponds to a stochastic process whose emitted symbols are consumed by a distortion-free decoder. This perspective reveals that many existing distortion-free schemes, including cyclic constructions, correspond to probabilistic deterministic finite automata (PDFAs) with relatively simple structure. The authors argue that watermark security is closely tied to the learnability of the underlying automaton and propose more expressive constructions based on probabilistic nondeterministic finite automata (PNFAs) to achieve stronger security guarantees.

Learning-theoretic hardness of structured distributions. The hardness results underlying automata-based watermarking draw on classical work in learning theory on the learnability of structured distributions. In particular, Kearns et al. (1994) study the problem of learning distributions generated by probabilistic automata and establish connections to cryptographic hardness assumptions such as learning parity with noise. These results provide a foundation for arguing that certain stochastic processes are not efficiently PAC learnable, even with access to evaluators.

3 Problem Derivation

We follow Wang & Shang (2025) to define watermarking a language model. Let $\mathcal{V}$ denote a token vocabulary where $\mathcal{V}^*$ is the set of all finite token sequences over the vocabulary.

Definition 3.1 (Unwatermarked Language Model)**.** An unwatermarked language model is a set of conditional distributions $p(\cdot \mid x)$ over $\mathcal{V}$, where $x \in \mathcal{V}^*$ is a prefix (e.g. a prompt). We generate text by sampling

$$y_t \sim p(\cdot \mid x_t), \qquad x_{t+1} = x_t \circ y_t,$$

where $\circ$ denotes concatenation.

Definition 3.2 (Watermarking Scheme)**.** A watermarking scheme for a language model consists of a tuple

$$\mathsf{WM} = (\mathsf{Gen}, \mathsf{Det}, \mathcal{K}),$$

where

- $K \leftarrow \mathcal{K}$ denotes a secret key sampled according to a specified key distribution.
- Gen_K is a randomized watermarked generator that, given a prompt $x \in \mathcal{V}^*$, induces a probability distribution over $\mathcal{V}^*$ and produces a random output sequence. The key K determines how sampling randomness is instantiated.
- $\mathsf{Det}_K : \mathcal{V}^* \to \{0, 1\}$ decides whether an input sequence was generated by Gen_K.

All watermarks are not created equal. Ideally, the watermark faithfully satisfies the properties from Bagchi et al. (2025): soundness, completeness, distortion-freeness, and robustness. We will define these properties formally.

Definition 3.3 (Soundness)**.** A watermarking scheme is δ-sound if for any prompt $x \in \mathcal{V}^*$ and any random sequence $Y \sim p(\cdot \mid x)$ generated by the unwatermarked language model,

$$\Pr[\mathsf{Det}_K(Y) = 1] \;\leq\; \delta.$$

Definition 3.4 (Completeness)**.** A watermarking scheme is δ-complete if for every prompt $x \in \mathcal{V}^*$,

$$\Pr[\mathsf{Det}_K(\mathsf{Gen}_K(x)) = 1] \;\geq\; 1 - \delta.$$

Definition 3.5 (Distortion-Freeness)**.** A watermarking scheme is distortion-free if for every prompt $x \in \mathcal{V}^*$ and every set $A \subseteq \mathcal{V}^*$,

$$\Pr[\mathsf{Gen}_K(x) \in A] \;=\; \Pr[Y \in A \mid Y \sim p(\cdot \mid x)]\,.$$

A watermarking scheme is ε-approximately distortion-free if for all prompts $x \in \mathcal{V}^*$,

$$\mathrm{D_{TV}}(\mathsf{Gen}_K(x),\; p(\cdot \mid x)) \;\leq\; \varepsilon,$$

where Kuditipudi et al. (2024) defines $\mathrm{D_{TV}}$ as the total variation distances.

Definition 3.6 (Robustness to Transformations)**.** Let $\mathcal{T}$ be a family of transformations $\tau : \mathcal{V}^* \to \mathcal{V}^*$. A watermarking scheme is δ-robust to $\mathcal{T}$ if for every prompt $x \in \mathcal{V}^*$ and every $\tau \in \mathcal{T}$,

$$\Pr[\mathsf{Det}_K(\tau(\mathsf{Gen}_K(x))) = 1] \;\geq\; 1 - \delta.$$

We contribute an additional desirable watermarking property: PAC security of the detector, which captures resistance to learning-based attacks. In this threat model, an adversary seeks to learn a hypothesis that approximates the watermark detector's decision rule (a function that reliably predicts whether a given text would be classified as watermarked) using polynomially many labeled examples obtained from detector queries or observed outputs. While distortion-freeness guarantees that individual watermarked samples are statistically indistinguishable from unwatermarked text, it does not prevent the detector itself from being learned from labeled data. PAC security rules out such attacks by ensuring that no efficient learner can approximate the detector with low error from polynomially many examples.

Definition 3.7 (PAC Security Against Detector Learning)**.** Let $\mathsf{WM} = (\mathsf{Gen}, \mathsf{Det}, \mathcal{K})$ be a watermarking scheme. Fix a distribution $\mathcal{D}$ over $\mathcal{V}^*$ and a key $K \leftarrow \mathcal{K}$. Let the target labeling function be

$$f^\star(x) = \mathsf{Det}_K(x).$$

We say that WM is (ε, δ)-PAC secure against detector learning with respect to $\mathcal{D}$ if for every learning algorithm $\mathcal{A}$ that is given access to $m = \mathrm{poly}(|K|, 1/\varepsilon, \log(1/\delta))$ i.i.d. labeled samples

$$\{(x_i, f^\star(x_i))\}_{i=1}^m, \qquad x_i \sim \mathcal{D},$$

the hypothesis $h \leftarrow \mathcal{A}$ satisfies

$$\Pr[\mathrm{err}_{\mathcal{D}}(h) \ \leq \ \varepsilon] \ \leq \ \delta.$$

4 Distortion-Free Watermarking

As a point of comparison, we formalize the red-green list watermarking paradigm introduced in Kirchenbauer et al. (2024), which biases the sampling distribution in favor of a subset of the vocabulary coined the "green list."

Definition 4.1 (Red-Green List Watermarking)**.** Let $\mathcal{V}$ be a vocabulary and let $p(\cdot \mid x)$ denote the base language model distribution given prefix $x \in \mathcal{V}^*$. A red-green watermarking scheme is parameterized by:

- a secret key K,
- a context length $r \geq 1$, and
- a hash function $h_K : \mathcal{V}^r \to \{0,1\}^{|\mathcal{V}|}$.

At generation step t, let $x_{t-r:t-1}$ denote the most recent r tokens. The hash $h_K(x_{t-r:t-1})$ induces a partition of the vocabulary into a *green set*

$$G_t = \{v \in \mathcal{V} : h_K(x_{t-r:t-1})_v = 1\}$$

and a *red set* $\mathcal{V} \setminus G_t$.

The watermarked generator samples the next token y_t from the biased distribution

$$p_K(y \mid x_t) \ \propto \ p(y \mid x_t) \cdot \exp\bigl(\gamma \cdot \mathbf{1}\{y \in G_t\}\bigr),$$

where $\delta > 0$ is a fixed bias parameter.

Given a candidate text $y = (y_1, \ldots, y_n)$, the detector recomputes the green sets $\{G_t\}$ using the same key K and applies a statistical hypothesis test to determine whether the empirical fraction of green tokens

$$\frac{1}{n} \sum_{t=1}^{n} \mathbf{1}\{y_t \in G_t\}$$

significantly exceeds its expectation under unwatermarked sampling.

We adopt the distortion-free watermarking framework of Kuditipudi et al. (2024) as our starting point. Their key insight is that watermarking can be achieved without changing the language model's output distribution by embedding a secret correlation into the *sampling randomness* rather than biasing token probabilities. The purpose of this section is to (i) restate the relevant components of this framework in our notation and (ii) define a cyclic-key specialization that will be the object of our learning-theoretic analysis in Section 5. All distributional and robustness guarantees in this section follow the methodology of Kuditipudi et al. (2024).

4.1 Distortion-Freeness

Let $V = \{1, \dots, N\}$ be a finite vocabulary and let $\Delta(V)$ denote the probability simplex over V. For $\mu \in \Delta(V)$ we write $\mu(i)$ for the mass on token i.

Definition 4.2 (Distortion-free decoder)**.** A (possibly randomized) decoder Γ is *distortion-free* if for every $\mu \in \Delta(V)$, the output distribution of $\Gamma(\mu)$ equals μ, i.e.,

$$\forall y \in V, \qquad \mathbb{P}[\Gamma(\mu) = y] = \mu(y).$$

Theorem 4.3 (Inverse Transform Sampling is distortion-free)**.** *Fix any permutation π of V. For any $\mu \in \Delta(V)$ define the cumulative sums*

$$C_k(\mu) \ := \ \sum_{j=1}^{k} \mu(\pi(j)), \qquad k = 1, \dots, N, \quad \textit{with } C_0(\mu) := 0.$$

Let $U \sim \mathrm{Unif}([0,1])$, and define the decoder

$$\Gamma_{\mathrm{ITS}}(\mu) \ := \ \pi(K) \quad \textit{where} \quad K := \min\{k \in \{1, \dots, N\} : U \le C_k(\mu)\}.$$

Then Γ_{ITS} is distortion-free: for every $\mu \in \Delta(V)$ and every $y \in V$,

$$\mathbb{P}\big[\Gamma_{\mathrm{ITS}}(\mu) = y\big] = \mu(y).$$

Proof. Fix $\mu \in \Delta(V)$ and $y \in V$. Let k_y be the unique index such that $\pi(k_y) = y$. By construction,

$$\{\Gamma_{\mathrm{ITS}}(\mu) = y\} \iff \{K = k_y\} \iff \{C_{k_y-1}(\mu) < U \le C_{k_y}(\mu)\}.$$

Since $U \sim \mathrm{Unif}([0,1])$, the probability of this event is the length of the interval:

$$\mathbb{P}\big[C_{k_y-1}(\mu) < U \le C_{k_y}(\mu)\big] = C_{k_y}(\mu) - C_{k_y-1}(\mu) = \mu(\pi(k_y)) = \mu(y).$$

This holds for all $y \in V$, hence Γ_{ITS} is distortion-free. □

Theorem 4.4 (Exponential-Minimum Sampling is distortion-free)**.** *Let $E_1, \dots, E_N$ be i.i.d.* $\mathrm{Exp}(1)$ *random variables. For any $\mu \in \Delta(V)$ define*

$$\Gamma_{\mathrm{EXP}}(\mu) \ := \ \arg \min_{i \in V:\, \mu(i) > 0} \frac{E_i}{\mu(i)},$$

with any deterministic tie-breaking rule (ties occur with probability 0*). Then Γ_{EXP} is distortion-free: for every $\mu \in \Delta(V)$ and every $y \in V$,*

$$\mathbb{P}\big[\Gamma_{\mathrm{EXP}}(\mu) = y\big] = \mu(y).$$

Proof. Fix $\mu \in \Delta(V)$ and $y \in V$ with $\mu(y) > 0$ (if $\mu(y) = 0$ the claim is trivial). Define the scaled variables

$$Z_i := \frac{E_i}{\mu(i)} \quad (i \in V,\ \mu(i) > 0).$$

For $t \ge 0$ and any i with $\mu(i) > 0$,

$$\mathbb{P}[Z_i > t] = \mathbb{P}[E_i > \mu(i)\, t] = e^{-\mu(i)t},$$

so $Z_i \sim \mathrm{Exp}(\mu(i))$, and the Z_i are independent.

We compute the probability that y attains the minimum:

$$\mathbb{P}[\Gamma_{\mathrm{EXP}}(\mu) = y] = \mathbb{P}[Z_y < Z_i \ \forall i \ne y]\,.$$

Using the law of total probability by conditioning on $Z_y = t$ and independence,

$$\begin{aligned}
\mathbb{P}[Z_y < Z_i \ \forall i \ne y] &= \int_0^\infty \mathbb{P}[Z_i > t \ \forall i \ne y \mid Z_y = t]\, f_{Z_y}(t)\, dt \\
&= \int_0^\infty \left(\prod_{i \ne y} \mathbb{P}[Z_i > t] \right) \cdot f_{Z_y}(t)\, dt \\
&= \int_0^\infty \left(\prod_{i \ne y} e^{-\mu(i)t} \right) \cdot \big(\mu(y) e^{-\mu(y)t}\big)\, dt \\
&= \int_0^\infty \mu(y)\, e^{-(\sum_{i \in V} \mu(i))t}\, dt.
\end{aligned}$$

Since $\sum_{i\in V}\mu(i)=1$, the integral equals

$$\mu(y)\int_0^\infty e^{-t}\,dt=\mu(y).$$

Thus $\mathbb{P}[\Gamma_{\mathrm{EXP}}(\mu)=y]=\mu(y)$ for all $y\in V$, proving distortion-freeness. □

If $U_i \sim \mathrm{Unif}(0,1)$ i.i.d. and $G_i := -\log(-\log U_i)$ are i.i.d. $\mathrm{Gumbel}(0,1)$, then $\arg\max_i\{\log\mu(i)+G_i\}$ has the same law as $\Gamma_{\mathrm{EXP}}(\mu)$ above. This is equivalent to the standard Gumbel-max trick described in Aaronson & Kirchner (2022).

4.2 Detection Guarantees

In this subsection, we formally state detectability guarantees for the two distortion-free sampling procedures introduced in Section 4.1: inverse transform sampling and exponential-minimum sampling. The results in this subsection are adapted from the analysis of Kuditipudi et al., and are restated here to make explicit the statistical power of alignment-based detectors used throughout this paper.

Let Gen_K denote a distortion-free generator parameterized by a secret key $K\in\Sigma^n$, and let Det_K be the alignment-based detector defined in Section 4.3. Let $x=(x_1,\dots,x_m)\in\mathcal{V}^m$ be a length-m text generated by Gen_K. We write $\alpha(x)$ for the watermark potential of the generated text, which measures the average strength of dependence between the sampled tokens and the underlying watermark randomness (see Kuditipudi et al. (2024) for a precise definition).

Theorem 4.5 (Detectability under inverse transform sampling)**.** *Suppose* Gen_K *implements distortion-free watermarking via inverse transform sampling. Then there exists a constant* $C>0$ *such that, for any text* x *of length* m *generated by* Gen_K*,*

$$\Pr\left[\mathsf{Score}_K(x)\ \le\ \min_{K'\neq K}\mathsf{Score}_{K'}(x)\right]\ \ge\ 1-2n\exp\bigl(-C\,m\,\alpha(x)^2\bigr)\,.$$

Theorem 4.6 (Detectability under exponential-minimum sampling)**.** *Suppose* Gen_K *implements distortion-free watermarking via exponential-minimum (Gumbel-max) sampling. Then there exists a universal constant* $C'>0$ *such that, for any text* x *of length* m *generated by* Gen_K*,*

$$\Pr\left[\mathsf{Score}_K(x)\ \le\ \min_{K'\neq K}\mathsf{Score}_{K'}(x)\right]\ \ge\ 1-2n\exp\bigl(-C'\,m\,\min\{\alpha(x),\alpha(x)^2\}\bigr)\,.$$

Proofs of the above theorems are non-trivial and are provided in Kuditipudi et al. (2024). That said, even if a watermarked is theoretically detectable, it still must also be robust to bounded deletions from an adversary.

4.3 Robustness Guarantees

Corollary 4.7 (Robust detectability under edit-distance perturbations)**.** *The detection guarantees of Theorems 4.5 and 4.6 continue to hold under bounded edit-distance transformations. Specifically, if* x' *is obtained from* x *by* e *edit operations, then watermark detection succeeds provided the additive increase in alignment score does not exceed the detection margin, as formalized in Proposition 4.9.*

To bolster robustness, we follow Kuditipudi et al. (2024) in the cyclic interpretation of the key. In practical settings, a watermarked text may be truncated and thus misaligned relative to the underlying randomness sequence used during generation. By reusing the key cyclically and allowing the detector to search over all offsets, the alignment procedure can recover the correct correspondence between text tokens and key symbols up to an unknown shift. This enables reliable detection even when the beginning of the generated sequence is missing or when insertions and deletions are present. Without cyclic reuse, detection would require exact positional synchronization between the text and the key, which is brittle under even very small edit-distance attacks.

Definition 4.8 (Cyclic Key)**.** Let Σ be a finite alphabet and let $L\in\mathbb{N}$. A *cyclic key* is a string

$$K=(K[1],K[2],\dots,K[L])\in\Sigma^L$$

interpreted modulo cyclic shift. That is, for any offset $s \in \{0, 1, \dots, L-1\}$ and any index $i \in \mathbb{N}$, we define

$$K[s+i] \triangleq K[((s+i-1) \bmod L) + 1].$$

The infinite key stream induced by K is the periodic extension

$$(K[1], K[2], \dots, K[L], K[1], K[2], \dots),$$

and any two keys that differ by a cyclic rotation are considered equivalent.

4.3.1 Alignment score

Let $x = (x_1, \dots, x_n) \in \mathcal{V}^n$ be a candidate text. Define an alignment cost between x and a length-n key-expansion of K under an unknown offset. Concretely, let $K_{1:n}^{(s)}$ denote the length-n cyclic expansion starting at offset s:

$$K_{1:n}^{(s)} = \big(K[s+1], K[s+2], \dots, K[s+n]\big) \in \Sigma^n.$$

We assume a compatibility score

$$\phi : \mathcal{V} \times \Sigma \to \mathbb{R}_{\geq 0}$$

that evaluates how consistent token x_i is with key symbol $K_i^{(s)}$.

To handle edits, we define an edit-distance-style dynamic program allowing insertions, deletions, and substitutions on the text. Let $c_{\text{ins}}, c_{\text{del}} \geq 0$ be insertion/deletion penalties. Define $D(i, j)$ as the minimum cost to align the prefix $(x_1, \dots, x_i)$ to the prefix $(K_1^{(s)}, \dots, K_j^{(s)})$:

$$D(0,0) = 0, \quad D(i,0) = i\, c_{\text{del}}, \quad D(0,j) = j\, c_{\text{ins}},$$

and for $i, j \geq 1$,

$$D(i,j) = \min\Big\{D(i-1,j) + c_{\text{del}},\ D(i,j-1) + c_{\text{ins}},\ D(i-1,j-1) + \phi(x_i, K_j^{(s)})\Big\}.$$

The alignment cost for a fixed offset s is $\text{Cost}_{K,s}(x) = D(n,n)$. Finally, define the cyclic alignment score

$$\text{Score}_K(x) = \min_{s \in \{1,\dots,L\}} \text{Cost}_{K,s}(x).$$

Given a key K and threshold $\tau \geq 0$, define the detector

$$\mathsf{Det}_K(x) = \mathbf{1}\{\text{Score}_K(x) \leq \tau\}.$$

Intuitively, watermarked text yields unusually small alignment cost under the correct key, whereas non-watermarked text aligns no better than random.

We denote a robustness guarantee: small edit-distance corruption increases alignment cost by at most an additive amount proportional to the number of edits, so detection persists under bounded corruption rates. This is the same edit-distance robustness mechanism used in Kuditipudi et al. (2024).

Proposition 4.9 (Robustness under edit operations)**.** *Fix K and consider any text $x \in \mathcal{V}^n$. Let x' be obtained from x by applying e edit operations (insertions, deletions, substitutions), with substitution cost upper bounded by $c_{\max}$ and insertion/deletion penalties bounded by $c_{\max}$. Then*

$$\text{Score}_K(x') \leq \text{Score}_K(x) + e\, c_{\max}.$$

Consequently, if $\mathsf{Det}_K(x) = 1$ and $e\, c_{\max} \leq \tau - \text{Score}_K(x)$, then $\mathsf{Det}_K(x') = 1$ as well.

Proof. Each edit operation can be simulated in the dynamic program by taking at most one additional insertion, deletion, or substitution transition, incurring cost at most $c_{\max}$. Therefore an alignment of x to the optimal key expansion can be converted to an alignment of x' to the same expansion with additional cost at most $e\, c_{\max}$. Minimizing over offsets s yields the stated inequality. □

The cyclic key reuse enhances robustness to cropping and edits by enabling the detector to search over offsets and align subsequences. However, the induced low-complexity cyclic structure also introduces a hypothesis class for learning under augmented access. Section 5 formalizes this via PAC learnability results for cyclic detectors, while Section 6 contrasts this with more complex key families based on probabilistic automata.

5 PAC Learnability of Cyclic Distortion-Free Watermarks

5.1 Learning Model and Target

We analyze learnability in a setting where an adversary has oracle access to the watermark detector. This is a standard supervised PAC model: there is an unknown distribution $\mathcal{D}$ over texts $x \in \mathcal{X}$, and an unknown target detector $\mathsf{Det}_{K^\star} \in \mathcal{H}_{\text{cyc}}$. The learner receives i.i.d. samples $x_1, \ldots, x_m \sim \mathcal{D}$ and labels $y_i = \mathsf{Det}_{K^\star}(x_i)$ obtained by querying the detector oracle. The goal is to output $\widehat{h}$ such that

$$\Pr_{x \sim \mathcal{D}} \left[\widehat{h}(x) \neq \mathsf{Det}_{K^\star}(x) \right] \leq \varepsilon$$

with probability at least $1 - \delta$ over the learner's sample.

5.2 Cyclic Alignment Detectors as a Finite Hypothesis Class

Fix an alphabet Σ and a key length $L \in \mathbb{N}$. A *cyclic key* is a string $K \in \Sigma^L$, interpreted modulo cyclic shift.

Let $\text{Score}_K : \mathcal{X} \to \mathbb{Z}$ be an integer-valued alignment score computable by a fixed dynamic program (e.g. edit-distance-style alignment) between the input text and the cyclic key K.[1] A threshold detector associated with (K, τ) outputs

$$h_{K,\tau}(x) = \mathbf{1}\{\text{Score}_K(x) \leq \tau\}.$$

We define the cyclic detector class

$$\mathcal{H}_{\text{cyc}} = \{h_{K,\tau} \ : \ K \in \Sigma^L, \ \tau \in \mathcal{T}\},$$

where $\mathcal{T}$ is a finite set of allowable thresholds. In particular, suppose there is a known bound $B \in \mathbb{N}$ such that for all $x \in \mathcal{X}$ and all $K \in \Sigma^L$,

$$\text{Score}_K(x) \in \{0, 1, \ldots, B\}. \tag{1}$$

Then it is without loss of generality to take $\mathcal{T} = \{0, 1, \ldots, B\}$, and hence

$$|\mathcal{H}_{\text{cyc}}| \leq |\Sigma|^L (B + 1). \tag{2}$$

5.3 Main Theorem and Proof

We prove that $\mathcal{H}_{\text{cyc}}$ is PAC learnable by empirical risk minimization (ERM). Because the true detector belongs to the class, ERM achieves realizable PAC learning.

Theorem 5.1 (PAC learnability of cyclic distortion-free detectors under detector queries)**.** *Assume equation 1 holds for some known B. Let $\mathcal{H}_{\text{cyc}}$ be defined as above, and suppose the target labels are generated by some $h^\star = h_{K^\star,\tau^\star} \in \mathcal{H}_{\text{cyc}}$ (realizable case). Then for any $\varepsilon, \delta \in (0, 1)$, the ERM learner over $\mathcal{H}_{\text{cyc}}$ is a PAC learner with sample complexity*

$$m \ \geq \ \frac{1}{\varepsilon} \left(\ln |\mathcal{H}_{\text{cyc}}| + \ln \frac{1}{\delta} \right) \ \leq \ \frac{1}{\varepsilon} \left(L \ln |\Sigma| + \ln(B + 1) + \ln \frac{1}{\delta} \right). \tag{3}$$

Specifically, with probability at least $1 - \delta$ over m i.i.d. samples from $\mathcal{D}$, ERM outputs $\widehat{h} \in \mathcal{H}_{\text{cyc}}$ satisfying

$$\Pr_{x \sim \mathcal{D}} [\widehat{h}(x) \neq h^\star(x)] \leq \varepsilon.$$

Proof. Define the true error of a hypothesis $h \in \mathcal{H}_{\text{cyc}}$ by

$$\text{err}_{\mathcal{D}}(h) = \Pr_{x \sim \mathcal{D}} [h(x) \neq h^\star(x)],$$

[1] The specific recurrence is irrelevant for PAC learnability; we only use that $\text{Score}_K(x)$ is well-defined and integer-valued.

and the *empirical error* on a sample $S = \{(x_i, y_i)\}_{i=1}^m$ by

$$\mathrm{err}_S(h) = \frac{1}{m} \sum_{i=1}^{m} \mathbf{1}\{h(x_i) \neq y_i\}, \quad \text{where } y_i = h^\star(x_i).$$

Let $\widehat{h}$ be any ERM solution:

$$\widehat{h} \in \arg\min_{h \in \mathcal{H}_{\mathrm{cyc}}} \mathrm{err}_S(h).$$

Because the setting is realizable, $h^\star \in \mathcal{H}_{\mathrm{cyc}}$ and hence $\mathrm{err}_S(h^\star) = 0$, which implies $\mathrm{err}_S(\widehat{h}) = 0$ as well.

We will show that with probability at least $1 - \delta$, every $h \in \mathcal{H}_{\mathrm{cyc}}$ with true error greater than ε must incur nonzero empirical error. This implies that any hypothesis with zero empirical error (in particular $\widehat{h}$) must have true error at most ε.

Fix any $h \in \mathcal{H}_{\mathrm{cyc}}$ with $\mathrm{err}_{\mathcal{D}}(h) > \varepsilon$. For each example $x_i \sim \mathcal{D}$, define the Bernoulli random variable

$$Z_i = \mathbf{1}\{h(x_i) \neq h^\star(x_i)\}.$$

Then $\mathbb{E}[Z_i] = \mathrm{err}_{\mathcal{D}}(h) > \varepsilon$, and

$$\mathrm{err}_S(h) = \frac{1}{m} \sum_{i=1}^{m} Z_i.$$

In particular, the event $\mathrm{err}_S(h) = 0$ is exactly the event $\sum_{i=1}^m Z_i = 0$. Since the Z_i are independent and $\Pr[Z_i = 1] = \mathrm{err}_{\mathcal{D}}(h)$, we have

$$\Pr[\mathrm{err}_S(h) = 0] = \Pr[Z_1 = 0, \ldots, Z_m = 0] = \prod_{i=1}^{m} (1 - \Pr[Z_i = 1]) = (1 - \mathrm{err}_{\mathcal{D}}(h))^m \leq (1 - \varepsilon)^m \leq e^{-\varepsilon m}.$$

Now apply a union bound over all hypotheses in $\mathcal{H}_{\mathrm{cyc}}$:

$$\Pr\left[\exists h \in \mathcal{H}_{\mathrm{cyc}} \text{ with } \mathrm{err}_{\mathcal{D}}(h) > \varepsilon \text{ and } \mathrm{err}_S(h) = 0\right] \leq |\mathcal{H}_{\mathrm{cyc}}| \cdot e^{-\varepsilon m}.$$

If $m \geq \frac{1}{\varepsilon}\left(\ln |\mathcal{H}_{\mathrm{cyc}}| + \ln \frac{1}{\delta}\right)$, then $|\mathcal{H}_{\mathrm{cyc}}| e^{-\varepsilon m} \leq \delta$. Hence, with probability at least $1 - \delta$, there does not exist any hypothesis with true error $> \varepsilon$ and empirical error 0. Since $\mathrm{err}_S(\widehat{h}) = 0$, it follows that $\mathrm{err}_{\mathcal{D}}(\widehat{h}) \leq \varepsilon$, proving the claim.

Finally, substituting equation 2 into the bound on m yields equation 3. □

Theorem 5.1 establishes *statistical* PAC learnability (sample complexity and generalization) of cyclic alignment-based detectors under detector-query access. The runtime of naive ERM may scale with $|\mathcal{H}_{\mathrm{cyc}}|$, which is exponential in L. Obtaining polynomial-time learnability requires additional algorithmic structure beyond finite-class arguments (e.g., efficient key reconstruction exploiting cyclic constraints), and is separable from PAC learnability.

6 Hardness of Probabilistic-Automata-Based Watermarks

Section 5 showed that cyclic distortion-free watermarking schemes induce detector classes of low complexity and are PAC learnable from polynomially many labeled examples. In this section, we show that this learnability is not inherent to distortion-free watermarking itself. Instead, it arises from the restricted structure of cyclic keys. By moving to more expressive stochastic processes (specifically, probabilistic automata) we obtain distortion-free watermarking schemes whose induced distributions are *not* efficiently PAC learnable under standard cryptographic assumptions.

6.1 Learning Model and Security Goal

We adopt a distribution-learning formulation consistent with classical results on learning probabilistic automata. Let $\mathcal{X}$ denote a finite output alphabet and let $\{D_K\}_{K \in \mathcal{K}}$ be a family of distributions over $\mathcal{X}^*$ induced by a watermarking scheme with key K.

A learner is given:

- i.i.d. samples $x_1, \dots, x_m \sim D_K$, and
- evaluator access to D_K, i.e., an oracle that returns $D_K(x)$ for any $x \in \mathcal{X}^*$.

The learner's goal is to output a hypothesis distribution $\widehat{D}$ such that

$$\mathrm{KL}(D_K \,\|\, \widehat{D}) \leq \varepsilon$$

with probability at least $1 - \delta$.

If an adversary can efficiently learn D_K in this sense, then it can approximate likelihoods under the watermark, emulate alignment-based detectors, and distinguish watermarked from unwatermarked text. Thus, PAC hardness of learning $\{D_K\}$ implies security against learning-based adversaries.

6.2 Hardness Assumption

Our hardness result relies on the standard assumption that learning parity with noise is computationally intractable.

Assumption 6.1 (Sparse Learning Parity with Noise (Sparse LPN))**.** *There is no polynomial-time algorithm that PAC learns k-sparse parity functions over $\{0,1\}^n$ under random classification noise, for $k = \omega(1)$.*

This assumption is widely used in cryptography and learning theory and underlies hardness results for learning structured distributions.(Blum et al., 1993)

6.3 Probabilistic-Automata-Based Watermarking

We now describe a class of watermarking schemes whose randomness is generated by a probabilistic nondeterministic finite automaton (PNFA). This construction follows the framework of Wang and Shang and generalizes cyclic-key schemes.

A PNFA $\mathcal{A}_K = (Q, \Sigma, \delta, \pi)$ consists of:

- a finite state set Q,
- an emission alphabet Σ,
- a probabilistic transition function δ, and
- a state-dependent emission distribution π.

The automaton generates an infinite sequence of symbols $Z_1, Z_2, \ldots \in \Sigma$ by stochastic transitions. These symbols are consumed by a distortion-free decoder (Section 4.1) to produce text whose marginal distribution exactly matches the base language model.

Crucially, the automaton's transition structure encodes a hidden parity function with additive noise. As a result, the induced output distribution D_K embeds a noisy parity instance into its stochastic dependencies, while remaining marginally indistinguishable from the unwatermarked model.

6.4 Main Hardness Result

We now state our main theorem.

Theorem 6.2 (PAC Hardness of Automata-Based Watermarks)**.** *Assuming Sparse LPN, there is no polynomial-time algorithm that PAC learns the distribution family $\{D_K\}$ induced by probabilistic-automata-based distortion-free watermarking schemes, even with evaluator access.*

Proof. Suppose for contradiction that there exists a polynomial-time learner $\mathcal{L}$ that PAC learns $\{D_K\}$ in KL divergence. We construct an algorithm $\mathcal{B}$ that uses $\mathcal{L}$ to learn sparse parity functions with noise, contradicting the Sparse LPN assumption.

Let $s \in \{0,1\}^n$ be a secret k-sparse parity vector. Using standard constructions from probabilistic automata theory, we define a PNFA $\mathcal{A}_s$ whose stochastic transitions encode the parity $\langle s, x \rangle$ with independent noise. The resulting emission process induces a distribution D_s over output sequences.

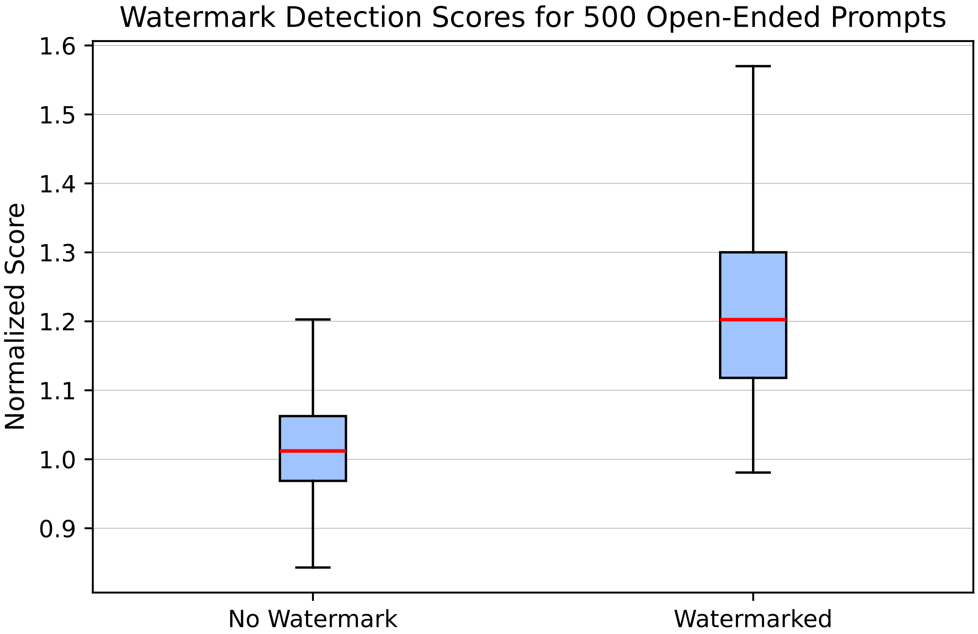

Figure 1: Distribution of normalized detection scores for unwatermarked as compared to watermarked text (500 open-ended prompts)

By assumption, $\mathcal{L}$ can, with polynomially many samples and evaluator queries, produce a hypothesis $\widehat{D}$ such that $\mathrm{KL}(D_s \,\|\, \widehat{D}) \leq \varepsilon$. From such a hypothesis, $\mathcal{B}$ can approximate likelihood ratios of carefully chosen events whose probabilities depend on the underlying parity function. Standard arguments show that this suffices to recover a hypothesis that predicts $\langle s, x \rangle$ with nontrivial advantage over random guessing, yielding a polynomial-time algorithm for Sparse LPN.

This contradicts the assumed hardness of Sparse LPN. Therefore, no such learner $\mathcal{L}$ exists. □

7 Experiments

In this section we present preliminary experiments evaluating the distortion-free watermarking scheme on the discrete diffusion model LLaDA (Nie et al., 2025). We defer modifications to the aforementioned watermarking schemes for discrete diffusion models (as opposed to an autoregressive model) to Bagchi et al. (2025). We will only show results for the the exponential minimum sampling scheme. In Figure 1 from my prior work in Bagchi et al. (2025), we show that the distortion-free watermarking scheme introduced in Kuditipudi et al. (2024); Aaronson & Kirchner (2022) is indeed detectable. In 2 and 3, we assess distortion-freeness and completeness respectively. We find that the watermark is indeed distortion-free (as perplexity does not increase) and it is detectable. Further experiments should implement the other optimizations from Kuditipudi et al. (2024) to improve results.

8 Conclusion

In this work, we studied the security of distortion-free language model watermarking through the lens of PAC learning. Although distortion-free schemes achieve information-theoretic invisibility by preserving the exact output distribution of the underlying language model, we showed that this guarantee alone does not protect against learning-based adversaries. Focusing on cyclic distortion-free watermarking schemes with alignment-based detectors, we proved that the induced detector class has low complexity and is PAC learnable from polynomially many labeled examples under natural adversary models, such as detector-query access. This result provides a principled explanation

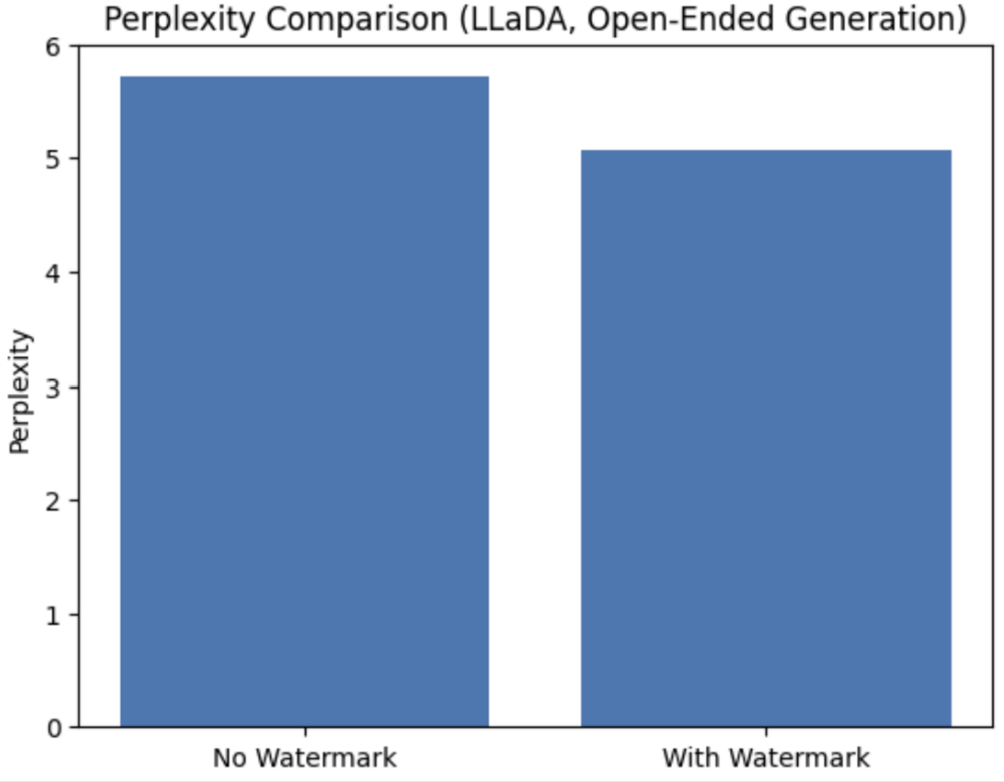

Figure 2: Perplexity comparison with and without watermark on LLaDA

for why cyclic watermarking schemes remain vulnerable despite strong statistical detectability and robustness guarantees.

Crucially, we demonstrated that this vulnerability is not inherent to distortion-free watermarking itself. By leveraging probabilistic-automata-based constructions and standard cryptographic hardness assumptions, we showed that more expressive watermarking schemes can induce distributions that are computationally hard to PAC learn, even with evaluator access. This establishes a sharp separation between cyclic and automata-based distortion-free watermarks and highlights learning complexity as a central determinant of watermark security. Our findings suggest that future watermarking designs should prioritize the unlearnability of the underlying stochastic process, positioning learning theory as a foundational tool for reasoning about robustness and security in language model watermarking

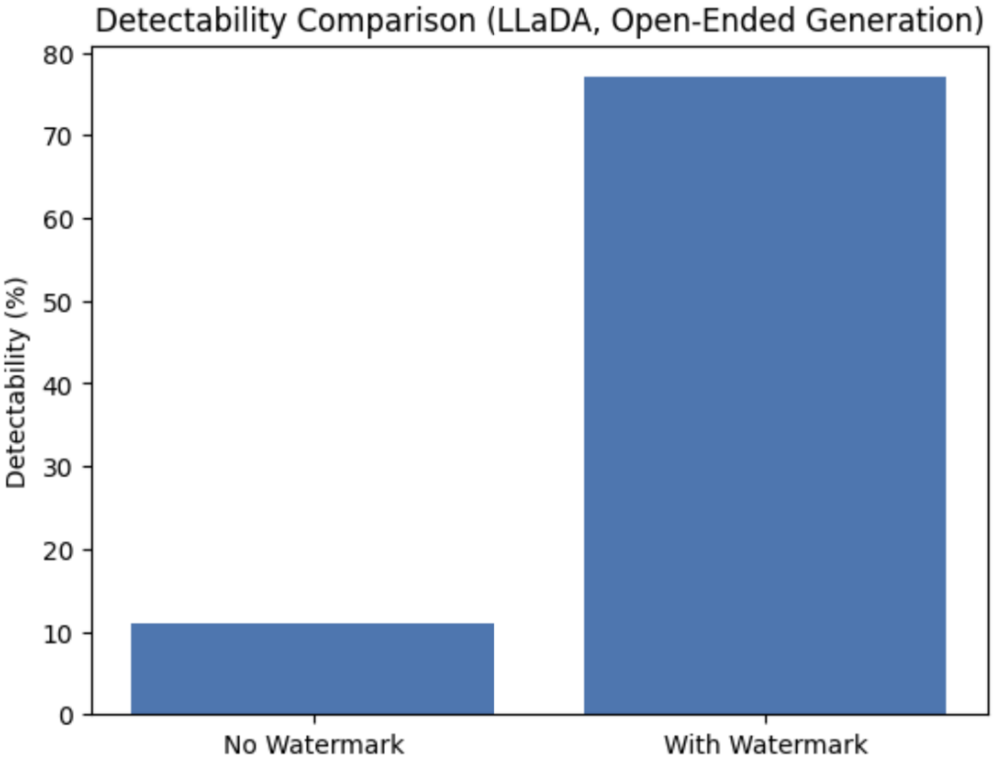

Figure 3: Completeness comparison with and without watermark on LLaDA

References

Scott Aaronson and Hendrik Kirchner. Watermarking gpt outputs. Lecture slides https://www.scottaaronson.com/talks/watermark.ppt, 2022. Accessed: 2025-10-13.

Avi Bagchi, Akhil Bhimaraju, Moulik Choraria, Daniel Alabi, and Lav R. Varshney. Watermarking discrete diffusion language models, 2025. URL `https://arxiv.org/abs/2511.02083`.

Avrim Blum, Merrick Furst, Michael Kearns, and Richard J. Lipton. Cryptographic primitives based on hard learning problems. In *Proceedings of the 13th Annual International Cryptology Conference (CRYPTO)*, pp. 278–291. Springer, 1993.

Michael Kearns, Yishay Mansour, Dana Ron, Ronitt Rubinfeld, Robert E. Schapire, and Linda Sellie. On the learnability of discrete distributions. In *Proceedings of the 26th 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, 2024. URL `https://arxiv.org/abs/2301.10226`.

Rohith Kuditipudi, John Thickstun, Tatsunori Hashimoto, and Percy Liang. Robust distortion-free watermarks for language models, 2024. URL `https://arxiv.org/abs/2307.15593`.

Shen Nie, Fengqi Zhu, Zebin You, Xiaolu Zhang, Jingyang Ou, Jun Hu, Jun Zhou, Yankai Lin, Ji-Rong Wen, and Chongxuan Li. Large language diffusion models, 2025. URL `https://arxiv.org/abs/2502.09992`.

Yangkun Wang and Jingbo Shang. Watermarks for language models via probabilistic automata, 2025. URL `https://arxiv.org/abs/2512.10185`.