

Markov Chain Lock Watermarking: Provably Secure Authentication for LLM Outputs¹

Chengheng Li Chen Kyuhee Kim
AI Safety Barcelona & EPFL MATS & EPFL
chengheng.lichen@epfl.ch kyuhee.kim@epfl.ch

With
Apart Research

Abstract

We present Markov Chain Lock (MCL) watermarking, a cryptographically secure framework for authenticating LLM outputs. MCL constrains token generation to follow a secret Markov chain over SHA-256 vocabulary partitions. Using doubly stochastic transition matrices, we prove four theoretical guarantees: (1) exponentially decaying false positive rates via Hoeffding bounds, (2) graceful degradation under adversarial modification with closed-form expected scores, (3) information-theoretic security without key access, and (4) bounded quality loss via KL divergence. Experiments on 173 Wikipedia prompts using Llama-3.2-3B demonstrate that the optimal 7-state soft cycle configuration achieves 100% detection, 0% FPR, and perplexity 4.20. Robustness testing confirms detection above 96% even with 30% word replacement. The framework enables $O(n)$ model-free detection, addressing EU AI Act Article 50 requirements. Code available at <https://github.com/ChenghengLi/MCLW>.

1. Introduction

The EU AI Act Article 50 [1] mandates machine-readable marking of AI-generated content, with enforcement beginning August 2026. Current approaches partition into passive detection [2, 3] and active watermarking [4]. Passive methods suffer high false positive rates and paraphrasing vulnerability [5]; active schemes degrade under token substitution and require probabilistic sampling incompatible with greedy decoding.

We introduce **Markov Chain Lock (MCL) watermarking**, constraining token generation to follow a secret finite-state Markov chain over cryptographic vocabulary partitions. Using doubly stochastic transition matrices (e.g., clockwork: $i \rightarrow (i + 1) \bmod S$, or soft cycle: $i \rightarrow \{i + 1, i + 2\}$), we create detectable state patterns verifiable only with the secret key (Figure 1).

Contributions. (1) Complete mathematical framework with cryptographic state assignment and clockwork topology; (2) four theorems establishing exponential false positive decay (Theorem 4.1), robustness bounds (Theorem 4.3), information-theoretic security (Theorem 4.4), and quality preservation (Theorem 4.5); (3) empirical validation on 173 Wikipedia concepts across 28 configurations; (4) practical algorithms for $O(N(V + T_{\text{LM}}))$ embedding and $O(n)$ model-free detection.

¹Research conducted at the Technical AI Governance Challenge, 2026. More information: <https://apartresearch.com/sprints/the-technical-ai-governance-challenge-2026-01-30-to-2026-02-01>

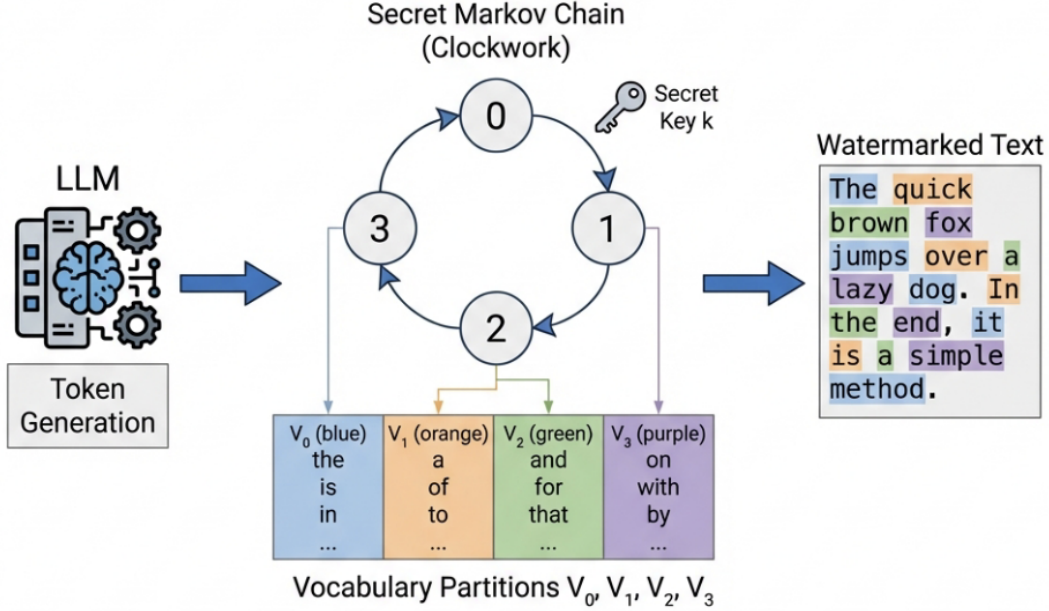


Figure 1: MCL watermarking overview. Each token is assigned to a state partition $\mathcal{V}_s$ via SHA-256 hashing. Generation follows a clockwork Markov chain ($0 \rightarrow 1 \rightarrow 2 \rightarrow 3 \rightarrow 0$), constraining each token to the next state’s partition. Detection verifies the hidden state sequence.

2. Related Work

Passive detection methods analyze statistical properties of generated text: DetectGPT [2] examines log-probability curvature, while GPTZero [3] combines perplexity and burstiness metrics. However, these approaches suffer high false positive rates on technical writing and fail under paraphrasing [5, 6]. Active watermarking embeds verifiable signals during generation. Kirchenbauer et al. [4] partition vocabulary into green/red lists via keyed hashing, but this degrades under paraphrasing and requires temperature sampling. Aaronson [7] proposes semantic-level embedding requiring expensive model queries, while Christ et al. [8] establish fundamental quality-security tradeoffs.

Chen et al. [9] propose Pattern-Mark, a Markov-chain-based watermarking framework for order-agnostic language models using probabilistic transitions and pattern-frequency detection. MCL targets autoregressive LLMs with deterministic transitions, enabling greedy decoding compatibility and model-free $O(n)$ detection via transition validity verification.

MCL advances prior work through: (1) deterministic embedding compatible with greedy decoding; (2) model-free $O(n)$ detection; (3) provable exponentially small FPR via Hoeffding bounds; and (4) explicit quality-security parameterization through tunable overlap ρ .

3. Methods

3.1 Cryptographic State Assignment

Let $\mathcal{V}$ denote the vocabulary with size $V = |\mathcal{V}|$, and let $H : \{0, 1\}^* \rightarrow \{0, 1\}^{256}$ denote SHA-256, a cryptographic hash function satisfying uniformity, one-wayness, and collision resistance.

Definition 3.1 (Cryptographic State Assignment). Given secret key $k \in \{0, 1\}^*$ and number of states $S \in \mathbb{N}$, the state assignment function is $\sigma_k : \mathcal{V} \rightarrow \mathcal{S}$ defined as $\sigma_k(t) = H(k||t) \bmod S$, where $||$ denotes concatenation and $\mathcal{S} = \{0, 1, \dots, S-1\}$. This induces vocabulary partition $\mathcal{V} = \bigsqcup_{s=0}^{S-1} \mathcal{V}_s$ where $\mathcal{V}_s = \{t \in \mathcal{V} : \sigma_k(t) = s\}$.

Lemma 3.2 (Partition Uniformity). *Under the random oracle model for H , for any state $s \in \mathcal{S}$: $\mathbb{E}[|\mathcal{V}_s|] = V/S$ and $\text{Var}[|\mathcal{V}_s|] = V(S-1)/S^2$.*

Proof. For each token t , the indicator $\mathbb{I}[\sigma_k(t) = s]$ is Bernoulli($1/S$) by hash uniformity. Thus $|\mathcal{V}_s| = \sum_{t \in \mathcal{V}} \mathbb{I}[\sigma_k(t) = s]$ has mean V/S and variance $V \cdot (1/S) \cdot (1 - 1/S) = V(S-1)/S^2$. $\square$

3.2 Transition Matrix Topologies

We consider transition matrices that are *doubly stochastic*, ensuring uniform stationary distribution, a property essential for the theoretical guarantees in Section 4.

Definition 3.3 (Doubly Stochastic Transition Matrix). A matrix $\mathbf{T} \in \mathbb{R}^{S \times S}$ is **doubly stochastic** if $\sum_j T_{ij} = \sum_i T_{ij} = 1$ for all i, j . For irreducible, aperiodic chains, the Ergodic Theorem guarantees convergence to the unique stationary distribution $\pi^* = (1/S, \dots, 1/S)$, yielding detection baseline $\mathbb{E}[\phi] = k/S$ where k is the number of valid successors per state.

We present two topologies satisfying this definition.

Example 3.4 (Clockwork Matrix). The clockwork matrix $\mathbf{T}_{ij}^{\text{clock}} = \mathbb{I}[j \equiv i+1 \pmod{S}]$ is a cyclic permutation with $k = 1$ successor per state. For $S = 4$:

$$\mathbf{T}^{\text{clock}} = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \end{pmatrix} \Rightarrow p_{\text{random}} = \frac{1}{S}$$

Example 3.5 (Soft Cycle Matrix). The soft cycle matrix allows transitions to two successors ($k = 2$):

$$\mathbf{T}_{ij}^{\text{soft}} = \begin{cases} 2/3 & j \equiv i+1 \pmod{S} \\ 1/3 & j \equiv i+2 \pmod{S} \\ 0 & \text{otherwise} \end{cases} \Rightarrow p_{\text{random}} = \frac{2}{S}$$

For $S = 4$:

$$\mathbf{T}^{\text{soft}} = \begin{pmatrix} 0 & 2/3 & 1/3 & 0 \\ 0 & 0 & 2/3 & 1/3 \\ 1/3 & 0 & 0 & 2/3 \\ 2/3 & 1/3 & 0 & 0 \end{pmatrix}$$

This doubles available vocabulary at each step while halving detection power relative to clockwork.

Note: For soft cycle with $S = 7$: $p_{\text{random}} = 2/7 \approx 0.29$.

3.3 Watermark Embedding and Detection

Definition 3.6 (MCL Constrained Generation). Given current state $s \in \mathcal{S}$, the MCL-constrained distribution over next tokens is:

$$P_{\text{MCL}}(t \mid s) = \begin{cases} P_{\mathcal{M}}(t)/Z_s & \text{if } t \in \mathcal{V}_{(s+1) \bmod S} \\ 0 & \text{otherwise} \end{cases}$$

where $Z_s = \sum_{t' \in \mathcal{V}_{(s+1) \bmod S}} P_{\mathcal{M}}(t')$ is the normalization constant.

Algorithm 1 MCL Watermark Embedding

Require: Prompt p , secret key k , states S , max tokens N

Ensure: Watermarked text

```
1: tokens  $\leftarrow$  TOKENIZE( $p$ )
2:  $s \leftarrow \sigma_k(\mathbf{tokens}[\text{last}])$  ▷ Initial state from prompt
3: for  $i = 1$  to  $N$  do
4:    $\ell \leftarrow \mathcal{M}.\text{FORWARD}(\mathbf{tokens})$  ▷ Get logits
5:    $s_{\text{next}} \leftarrow (s + 1) \bmod S$  ▷ Clockwork: next state
6:   for each token  $t \in \mathcal{V}$  do
7:     if  $\sigma_k(t) \neq s_{\text{next}}$  then
8:        $\ell[t] \leftarrow -\infty$  ▷ Mask disallowed tokens
9:     end if
10:  end for
11:   $t_{\text{next}} \leftarrow \arg \max(\ell)$  ▷ Greedy selection
12:  tokens.APPEND( $t_{\text{next}}$ );  $s \leftarrow s_{\text{next}}$ 
13:  if  $t_{\text{next}} = \text{EOS}$  then: break
14:  end if
15: end for
16: return DECODE( $\mathbf{tokens}[\text{len}(p) : ]$ )
```

Definition 3.7 (Fingerprint Score). For token sequence $\mathbf{t} = (t_1, \dots, t_n)$, the fingerprint score is:

$$\phi(\mathbf{t}) = \frac{1}{n-1} \sum_{i=1}^{n-1} \mathbf{T}_{\sigma_k(t_i), \sigma_k(t_{i+1})}^{\text{clock}}$$

Text is classified as watermarked if $\phi(\mathbf{t}) > \tau$ for detection threshold $\tau \in (1/S, 1)$.

3.4 Soft Partitions

Hard partitions restrict each token to exactly one state, which may limit vocabulary choices and degrade generation quality. Soft partitions allow controlled overlap between states.

Definition 3.8 (Soft State Membership). Given overlap ratio $\rho \in [0, 1]$, the soft membership assigns each token to multiple states: $\Sigma_k(t) = \{s \in \mathcal{S} : H_s(k||t) < \theta_\rho\}$ where $\theta_\rho = (1/S + \rho(1 - 1/S)) \cdot 2^{256}$.

The expected tokens per state is $V(1/S + \rho(S-1)/S)$. For $S = 7$: $\rho = 0$ yields 14.3% vocabulary per state, while $\rho = 0.15$ yields approximately 27%. Higher overlap improves quality but weakens detection, as demonstrated in our experiments.

3.5 Algorithms

Algorithm 1 has complexity $O(N(V + T_{\text{LM}}))$ where T_{LM} is the LLM forward pass time. Algorithm 2 requires only $O(n)$ hash operations with **no model access**, enabling lightweight third-party verification.

Algorithm 2 MCL Watermark Detection

Require: Text x , secret key k , states S , threshold τ

Ensure: ($is_watermarked, score, p_value$)

```
1: tokens  $\leftarrow$  TOKENIZE( $x$ );  $n \leftarrow |\mathbf{tokens}|$ ;  $c \leftarrow 0$ 
2: for  $i = 1$  to  $n - 1$  do
3:    $s_i \leftarrow \sigma_k(\mathbf{tokens}[i])$ ;  $s_{i+1} \leftarrow \sigma_k(\mathbf{tokens}[i + 1])$ 
4:   if  $s_{i+1} \equiv s_i + 1 \pmod{S}$  then:  $c \leftarrow c + 1$ 
5:   end if ▷ Valid clockwork transition
6: end for
7:  $score \leftarrow c/(n - 1)$ 
8:  $z \leftarrow (score - 1/S) / \sqrt{(1/S)(1 - 1/S)/(n - 1)}$  ▷ Z-score
9: return ( $score > \tau, score, 1 - \Phi(z)$ )
```

4. Theoretical Results

Theorem 4.1 (Detection Bounds for Doubly Stochastic Chains). *Let $\mathbf{T}$ be an irreducible doubly stochastic transition matrix with k non-zero entries per row (average). Let $\mathbf{t}^w$ be a watermarked sequence of length n and $\mathbf{t}^r$ a random sequence. For threshold $\tau \in (k/S, 1)$:*

- (i) *Watermarked:* $\phi(\mathbf{t}^w) = 1$ (all transitions valid by construction)
- (ii) *Random baseline:* $\mathbb{E}[\phi(\mathbf{t}^r)] = k/S$ (uniform state distribution)
- (iii) *False negative rate:* $FNR = 0$ ($\phi = 1 > \tau$ always)
- (iv) *False positive rate:* $FPR \leq \exp(-2(n-1)(\tau - k/S)^2)$ (Hoeffding bound)

Special cases: Clockwork ($k=1$): $p_{random} = 1/S$. Soft cycle ($k=2$): $p_{random} = 2/S$.

Proof. (i) Algorithm 1 enforces that each token's state is a valid successor in $\mathbf{T}$, so $\mathbf{T}_{\sigma_k(t_i), \sigma_k(t_{i+1})} > 0$ for all transitions, giving $\phi(\mathbf{t}^w) = 1$. (ii) For random text, each state $\sigma_k(t)$ is uniformly distributed over $\mathcal{S}$ by Lemma 3.2. Since $\mathbf{T}$ is doubly stochastic (Definition 3.3), the stationary distribution is uniform: $\pi^* = (1/S, \dots, 1/S)$. With k non-zero entries per row, $P(\text{valid transition}) = k/S$. (iii) Since $\phi(\mathbf{t}^w) = 1 > \tau$, $FNR = 0$. (iv) Let $Z_i = \mathbf{1}[\mathbf{T}_{\sigma_k(t_i), \sigma_k(t_{i+1})} > 0] \in \{0, 1\}$ indicate valid transitions. For random text, Z_i are i.i.d. Bernoulli(k/S) by the doubly stochastic property. By Hoeffding's inequality:

$$P(\phi(\mathbf{t}^r) \geq \tau) = P\left(\frac{1}{n-1} \sum_{i=1}^{n-1} Z_i \geq \tau\right) \leq \exp\left(-2(n-1) \left(\tau - \frac{k}{S}\right)^2\right)$$

□

Corollary 4.2 (Practical FPR Bounds). *For soft cycle ($k=2$), $S = 7$, $\tau = 0.5$: $FPR \leq 1.1 \times 10^{-4}$ at $n = 100$ tokens; $FPR \leq 1.2 \times 10^{-8}$ at $n = 200$ tokens.*

Theorem 4.3 (Adversarial Robustness). *For a doubly stochastic chain with k valid successors per state, if an adversary modifies fraction δ of tokens uniformly at random:*

$$\mathbb{E}[\phi(\mathbf{t}^{mod})] = (1 - \delta)^2 + \frac{k \cdot \delta(2 - \delta)}{S}$$

Proof. Consider consecutive pair (t_i, t_{i+1}) with modification indicator $M_i \sim \text{Bernoulli}(\delta)$. Four cases: (1) neither modified (prob. $(1-\delta)^2$): contributes 1; (2) only t_{i+1} modified (prob. $(1-\delta)\delta$):

contributes k/S ; (3) only t_i modified (prob. $\delta(1-\delta)$): contributes k/S ; (4) both modified (prob. δ^2): contributes k/S . Summing: $(1-\delta)^2 + [2\delta(1-\delta) + \delta^2] \cdot k/S = (1-\delta)^2 + \delta(2-\delta) \cdot k/S$. $\square$

Example: For soft cycle ($k=2$), $S = 7$, $\tau = 0.5$: at $\delta = 0.3$, $\mathbb{E}[\phi] = 0.49 + 0.51 \cdot 2/7 \approx 0.64 > 0.5$ (detectable); at $\delta = 0.5$, $\mathbb{E}[\phi] = 0.25 + 0.75 \cdot 2/7 \approx 0.46 < 0.5$ (undetectable). Experimental validation is presented in Section 6.3.

Theorem 4.4 (Information-Theoretic Security). *Under the random oracle model, without secret key k , any polynomial-time adversary has negligible advantage:*

$$\text{Adv}_{\mathcal{A}} = |P[\mathcal{A}(\mathbf{t}^w) = 1] - P[\mathcal{A}(\mathbf{t}^r) = 1]| \leq \text{negl}(256)$$

Proof. The state assignment $\sigma_k(t) = H(k\|t) \bmod S$ is a pseudorandom function under the random oracle model. Without k , the adversary observes only the token sequence $\mathbf{t}$, not the state sequence $(s_1, \dots, s_n)$. Watermarked text has clockwork states; random text has i.i.d. uniform states. Distinguishing these from tokens alone requires inverting H , which has probability $\leq 2^{-256}$. $\square$

Theorem 4.5 (Quality Bound). *The KL divergence between MCL-constrained and unconstrained distributions satisfies:*

$$D_{KL}(P_{MCL} \| P_{\mathcal{M}}) = \log(1/Z_s) \leq \log S$$

where $Z_s = P_{\mathcal{M}}(\mathcal{V}_{(s+1) \bmod S})$ is the probability mass of the target partition.

Proof. MCL distribution: $P_{MCL}(t | s) = P_{\mathcal{M}}(t)/Z_s$ for $t \in \mathcal{V}_{(s+1) \bmod S}$, else 0. KL divergence:

$$D_{KL} = \sum_{t \in \mathcal{V}_{(s+1) \bmod S}} \frac{P_{\mathcal{M}}(t)}{Z_s} \log \frac{1}{Z_s} = \log(1/Z_s)$$

By Lemma 3.2, $|\mathcal{V}_{(s+1) \bmod S}| \approx V/S$, so $Z_s \geq 1/S$, giving $D_{KL} \leq \log S$. $\square$

Implication: Quality degradation is logarithmic in states. For $S = 7$: $D_{KL} \leq \log 7 \approx 1.95$ nats.

5. Experimental Design

To empirically validate our theoretical framework, we conduct comprehensive experiments across diverse textual domains and watermarking configurations. This section details the dataset construction, text generation methodology, and evaluation criteria.

5.1 Dataset Construction

We construct an evaluation corpus comprising 173 concepts selected from the most frequently accessed Wikipedia articles, ensuring both topical diversity and practical relevance. The corpus systematically spans six semantic categories: *notable figures* (e.g., Albert Einstein, William Shakespeare, Elon Musk), *historical events* (e.g., World War II, French Revolution, the Renaissance), *scientific phenomena* (e.g., quantum mechanics, DNA replication, black holes), *geographical entities* (e.g., Mount Everest, Amazon River, Antarctica), *technological concepts* (e.g., artificial intelligence, cryptocurrency, neural networks), and *cultural references* (e.g., Star Wars, Harry Potter, Game of Thrones).

For each concept c , we construct a standardized prompt p_c of the form: “*Explain [concept] in a comprehensive way.*” This formulation elicits informative, multi-sentence responses suitable for watermark embedding and detection analysis.

5.2 Text Generation Protocol

We generate paired samples for each concept to enable controlled comparison between watermarked and baseline conditions:

- **Watermarked generation.** Text is produced using Algorithm 1 with **soft cycle** transition enforcement (Example 3.5). At each decoding step, the model’s output distribution is masked to permit only tokens belonging to valid successor partitions $\mathcal{V}_{(s+1) \bmod S}$ or $\mathcal{V}_{(s+2) \bmod S}$, followed by sampling from the masked distribution. This procedure guarantees $\phi(\mathbf{t}) \approx 1.0$ for all generated sequences.
- **Baseline generation.** Non-watermarked text is produced using standard nucleus sampling with temperature $T = 0.7$ and cumulative probability threshold $p = 0.9$, without any vocabulary constraints. These samples establish the null distribution for false positive rate estimation.

All experiments employ Llama-3.2-3B-Instruct as the base language model, with a maximum generation length of 200 tokens. We systematically evaluate $7 \times 4 = 28$ configurations spanning state counts $S \in \{2, 4, 5, 7, 9, 11, 15\}$ and overlap ratios $\rho \in \{0\%, 5\%, 10\%, 15\%\}$. Reproducibility is ensured through fixed random seed (42) and cryptographic key (`curated_wiki_dataset_2024`).

5.3 Evaluation Metrics

We quantify watermarking performance along four complementary dimensions:

- **Fingerprint Score** $\phi(\mathbf{t}) \in [0, 1]$: The proportion of consecutive token pairs exhibiting valid transitions according to the transition matrix (Definition 3.7). For watermarked text using soft cycle, $\phi \approx 1.0$ by construction; for random text, $\mathbb{E}[\phi] = 2/S$ (Theorem 4.1).
- **True Positive Rate (TPR)**: The fraction of watermarked samples correctly identified, computed as $\text{TPR} = |\{\mathbf{t}^w : \phi(\mathbf{t}^w) > \tau\}| / |\{\mathbf{t}^w\}|$ with threshold $\tau = 0.5$.
- **False Positive Rate (FPR)**: The fraction of non-watermarked samples erroneously classified as watermarked, representing the Type I error rate critical for deployment reliability.
- **Perplexity (PPL)**: The exponential of cross-entropy loss, $\text{PPL} = \exp\left(-\frac{1}{n} \sum_{i=1}^n \log P_{\mathcal{M}}(t_i | t_{<i})\right)$, measuring linguistic fluency. Lower values indicate higher generation quality.

5.4 Robustness Evaluation Protocol

To evaluate robustness against adversarial attacks, we perform controlled word replacement experiments on the optimal configuration ($S = 7$, soft cycle, $\rho = 0\%$). For each modification rate $\delta \in \{10\%, 20\%, 30\%, 40\%, 50\%, 60\%\}$, we randomly select δ fraction of word positions in each watermarked text and replace them with the word “masked”. The modified text is re-tokenized using Llama-3.2’s tokenizer, and detection scores are computed using Algorithm 2. Table 2 reports the observed mean fingerprint score ϕ across 173 samples, the theoretical expectation $\mathbb{E}[\phi]$ from Theorem 4.3, and the detection rate (fraction with $\phi > \tau = 0.5$).

6. Experimental Results

6.1 Main Results

Table 1 presents the top-performing configurations. All 0% overlap configurations with $S \geq 5$ achieve perfect detection (100% TPR, 0% FPR), confirming Theorem 4.1. The optimal $S = 7$ configuration balances detection with minimal perplexity (4.20). Complete results for all 28 configurations are provided in Appendix A.

Table 1: Top configurations (0% overlap): Score gap = WM avg – Non-WM avg

States	WM Avg	Non-WM Avg	Gap	Det%	FPR%	PPL
5	0.991	0.402	0.59	100	0	4.98
7	0.991	0.290	0.70	100	0	4.20
9	0.984	0.221	0.76	100	0	5.37
11	0.988	0.197	0.79	100	0	4.61
15	0.988	0.138	0.85	100	0	5.11

6.2 Key Observations

Our experiments reveal four critical findings. **First**, configurations with $S \geq 5$ and $\rho = 0\%$ achieve *perfect separability*: the minimum watermarked score exceeds the maximum non-watermarked score, enabling simultaneous 0% FPR and 0% FNR at threshold $\tau = 0.5$, directly confirming Theorem 4.1. **Second**, the optimal configuration is $S = 7$ with $\rho = 0\%$, achieving 100% TPR, 0% FPR, lowest perplexity (4.20), and strong separation (gap = 0.70), the Pareto-optimal choice consistent with Theorem 4.5. **Third**, low state counts fail: $S = 2$ yields 100% FPR (all sequences trivially alternate), while $S = 4$ suffers 33.5% FPR due to insufficient separation. **Fourth**, overlap catastrophically degrades detection: for $S = 7$, TPR drops from 100% ($\rho = 0\%$) to 92% ($\rho = 5\%$), 35% ($\rho = 10\%$), and 13% ($\rho = 15\%$).

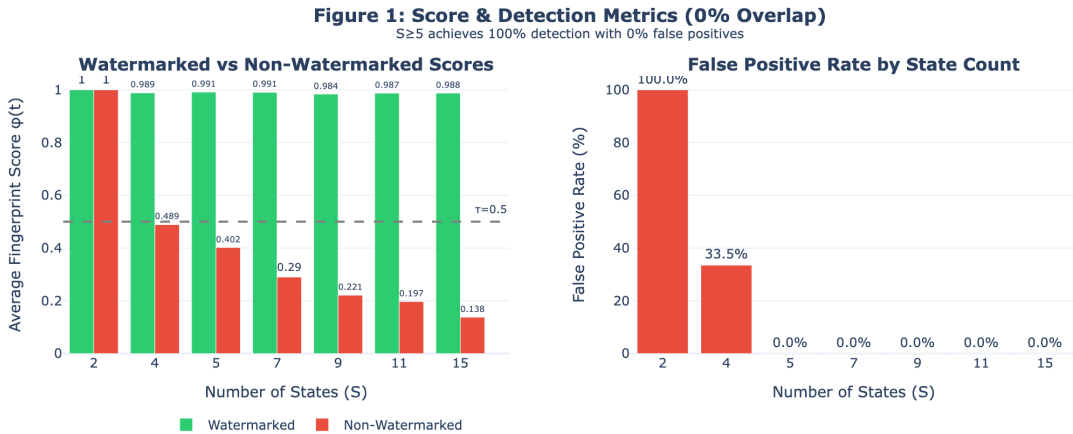


Figure 2: Detection metrics (0% overlap, soft cycle). **Left:** Watermarked text achieves $\phi \approx 1.0$; non-watermarked follows $2/S$ baseline. **Right:** FPR by state count; $S \leq 4$ yields unacceptable false positives.

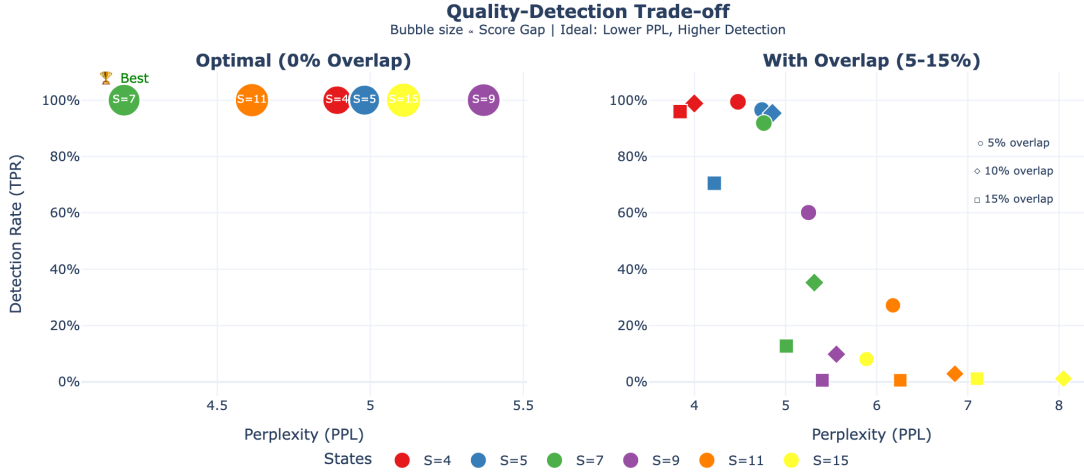


Figure 3: Quality-detection trade-off (bubble size $\propto$ score gap). $S = 7$ at 0% overlap minimizes PPL while achieving 100% detection.

6.3 Robustness Analysis

Table 2 presents the robustness validation results following the protocol in Section 5.4. The baseline (unmodified text) achieves $\phi = 0.991$ with 100% detection as shown in Table 1.

Table 2: Robustness to word replacement (soft cycle, $S = 7$, $n = 173$).

δ	Observed ϕ	Std Dev	$\mathbb{E}[\phi]$ (Thm. 4.3)	Detection Rate
0%	0.991	0.014	1.000	100%
10%	0.839	0.068	0.864	99.4%
20%	0.738	0.072	0.743	99.4%
30%	0.642	0.073	0.636	96.5%
40%	0.543	0.080	0.543	76.3%
50%	0.455	0.084	0.464	17.9%
60%	0.365	0.084	0.400	5.8%

Observed scores match Theorem 4.3 within 1 to 6%. Detection remains above 96% for $\delta \leq 30\%$. The critical threshold is $\delta \approx 45\%$ where $\mathbb{E}[\phi] = \tau = 0.5$. Beyond this point, an adversary must corrupt nearly half the text to evade detection, which would likely render the content incoherent.

7. Discussion

MCL watermarking addresses EU AI Act Article 50 [1] requirements for machine-readable AI content marking. The model-free detection (Algorithm 2) enables third-party auditing without LLM access, supporting transparency frameworks [10], while provable guarantees (Theorems 4.1 to 4.5) provide legally defensible authentication with quantifiable error rates.

Limitations. (1) Detection requires the same tokenizer used during generation; (2) at least 5 states are needed for reliable detection ($S = 2$ is unusable); (3) overlap $\rho \geq 10\%$ severely degrades detection; (4) perplexity increases with states, though $S = 7$ achieves reasonable PPL = 4.20.

Future Work. A promising direction is *sparse watermarking*: introducing a “full vocabulary” state that allows unrestricted token selection, watermarking only every N -th token (e.g., $N = 10$). This preserves generation quality while maintaining theoretical guarantees, since any doubly stochastic transition matrix ensures uniform stationary distribution. Other directions include adaptive overlap selection based on token probability mass, C2PA metadata integration for end-to-end provenance, robustness against learned adversarial attacks, and multi-model coordination for heterogeneous deployments. Preliminary experiments on Mistral-7B and GPT-2-XL (Appendix E) confirm consistent behavior across model architectures. We hypothesize that larger vocabulary sizes reduce perplexity under partitioning, as each partition contains more candidate tokens.

8. Conclusion

We established complete mathematical foundations for Markov Chain Lock watermarking using doubly stochastic transition matrices over SHA-256-based vocabulary partitions. Our theoretical results generalize to any irreducible doubly stochastic chain with k valid successors per state, providing: (1) exponentially decaying false positive rates bounded by Hoeffding’s inequality (Theorem 4.1), (2) quantified robustness under adversarial token modification with formula $\mathbb{E}[\phi] = (1 - \delta)^2 + k\delta(2 - \delta)/S$ (Theorem 4.3), (3) information-theoretic security under the random oracle model (Theorem 4.4), and (4) logarithmic quality degradation in state count (Theorem 4.5). Experiments on 173 Wikipedia concepts across 28 configurations demonstrate that 7-state soft cycle ($k = 2$) with 0% overlap is optimal, achieving 100% detection, 0% FPR, and perplexity 4.20. Robustness validation confirms detection remains above 96% even with 30% word modification. The framework provides $O(n)$ model-free detection, directly addressing EU AI Act Article 50 requirements.

LLM Usage Statement

Claude (Anthropic) assisted with LaTeX formatting. All proofs, algorithms, and experiments independently developed and verified by the authors.

Code and Data

Repository: <https://github.com/ChenghengLi/MCLW>. **Model:** Llama-3.2-3B-Instruct. **Dataset:** 173 curated Wikipedia concepts, seed 42, secret key `curated_wiki_dataset_2024`.

Author Contributions

Chengheng Li: Theoretical framework design, mathematical proofs, algorithm implementation, experiments, manuscript writing. **Kyuhee Kim:** Experimental design, theoretical review, AI governance analysis.

Acknowledgments

We thank Apart Research and the Technical AI Governance Challenge organizers for hosting this event.

References

- [1] European Parliament and Council. EU AI Act (Regulation 2024/1689), 2024. URL <https://eur-lex.europa.eu/eli/reg/2024/1689>. Official Journal of the European Union.
- [2] Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D. Manning, and Chelsea Finn. DetectGPT: Zero-shot machine-generated text detection using probability curvature. In *International Conference on Machine Learning (ICML)*, pages 24950–24962, 2023.
- [3] Edward Tian. GPTZero: An AI plagiarism detection tool, 2023. URL <https://gptzero.me>. Software tool.
- [4] 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 (ICML)*, pages 17061–17084, 2023.
- [5] Kalpesh Krishna, Yixiao Song, Marzena Karpinska, John Wieting, and Mohit Iyyer. Paraphrasing evades detectors of AI-generated text, but retrieval is an effective defense. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 13287–13300, 2023.
- [6] Vinu Sankar Sadasivan, Aounon Kumar, Sriram Balasubramanian, Wenxiao Wang, and Soheil Feizi. Can AI-generated text be reliably detected? *arXiv preprint arXiv:2303.11156*, 2023.
- [7] Scott Aaronson. Watermarking GPT outputs. 2023. URL <https://scottaaronson.blog/?p=6823>. Blog post.
- [8] Miranda Christ, Sam Gunn, and Or Zamir. Undetectable watermarks for language models, 2023. URL <https://arxiv.org/abs/2306.09194>.
- [9] Ruibo Chen, Yihan Wu, Yanshuo Chen, Chenxi Liu, Junfeng Guo, and Heng Huang. A watermark for order-agnostic language models, 2024. URL <https://arxiv.org/abs/2410.13805>.
- [10] OECD. Hiroshima AI process international code of conduct for organizations developing advanced AI systems, 2024. URL <https://oecd.ai/en/g7-reporting-framework>. G7 Hiroshima AI Process.

A. Complete Experimental Results

Table 3 presents complete results for all 28 configurations. Verdict classification: **Excellent** (gap > 0.5 , 100% detection, 0% FPR); **Good** (gap > 0.3 , $>90\%$ detection); **Marginal** (gap > 0.15); **Poor** (gap < 0.15 or high FPR).

Table 3: Complete results for all 28 configurations

Config	S	ρ	Non-WM	WM	Gap	Det%	FPR%	PPL
states2_overlap0pct	2	0%	1.000	1.000	0.00	100	100	1.29
states2_overlap5pct	2	5%	1.000	1.000	0.00	100	100	1.29
states2_overlap10pct	2	10%	1.000	1.000	0.00	100	100	1.29
states2_overlap15pct	2	15%	1.000	1.000	0.00	100	100	1.29
states4_overlap0pct	4	0%	0.489	0.989	0.50	100	33.5	4.89
states4_overlap5pct	4	5%	0.489	0.842	0.35	99.4	33.5	4.48
states4_overlap10pct	4	10%	0.489	0.696	0.21	98.8	33.5	4.00
states4_overlap15pct	4	15%	0.489	0.626	0.14	96.0	33.5	3.84
states5_overlap0pct	5	0%	0.402	0.991	0.59	100	0	4.98
states5_overlap5pct	5	5%	0.402	0.726	0.32	96.5	0	4.74
states5_overlap10pct	5	10%	0.402	0.639	0.24	95.4	0	4.86
states5_overlap15pct	5	15%	0.402	0.542	0.14	70.5	0	4.22
states7_overlap0pct	7	0%	0.290	0.991	0.70	100	0	4.20
states7_overlap5pct	7	5%	0.290	0.643	0.35	91.9	0	4.76
states7_overlap10pct	7	10%	0.290	0.477	0.19	35.3	0	5.32
states7_overlap15pct	7	15%	0.290	0.414	0.12	12.7	0	5.01
states9_overlap0pct	9	0%	0.221	0.984	0.76	100	0	5.37
states9_overlap5pct	9	5%	0.221	0.531	0.31	60.1	0	5.25
states9_overlap10pct	9	10%	0.221	0.379	0.16	9.8	0	5.56
states9_overlap15pct	9	15%	0.221	0.307	0.09	0.6	0	5.40
states11_overlap0pct	11	0%	0.197	0.987	0.79	100	0	4.61
states11_overlap5pct	11	5%	0.197	0.431	0.23	27.2	0	6.18
states11_overlap10pct	11	10%	0.197	0.315	0.12	2.9	0	6.86
states11_overlap15pct	11	15%	0.197	0.255	0.06	0.6	0	6.26
states15_overlap0pct	15	0%	0.138	0.988	0.85	100	0	5.11
states15_overlap5pct	15	5%	0.138	0.347	0.21	8.1	0	5.89
states15_overlap10pct	15	10%	0.138	0.257	0.12	1.2	0	8.05
states15_overlap15pct	15	15%	0.138	0.203	0.07	1.2	0	7.10

B. Separability and Overlap Analysis

Table 4 demonstrates perfect separability: for $S \geq 5$ with $\rho = 0\%$, the minimum watermarked score exceeds the maximum non-watermarked score, enabling zero-error detection.

Table 4: Score range separability and overlap impact

S	Separability (0% overlap)				Detection Rate by Overlap			
	WM Min	Non-WM Max	Gap	Sep?	0%	5%	10%	15%
5	0.607	0.500	+0.11	Yes	100%	96.5%	95.4%	70.5%
7	0.707	0.400	+0.31	Yes	100%	91.9%	35.3%	12.7%
9	0.653	0.333	+0.32	Yes	100%	60.1%	9.8%	0.6%
11	0.720	0.313	+0.41	Yes	100%	27.2%	2.9%	0.6%
15	0.787	0.247	+0.54	Yes	100%	8.1%	1.2%	1.2%

C. Supplementary Visualizations

This section provides additional visualizations supporting the main experimental findings.

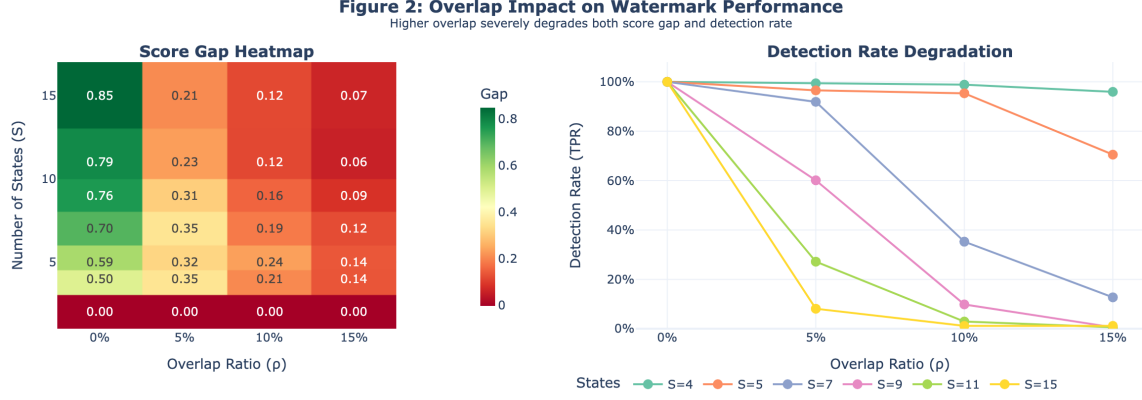


Figure 4: Overlap impact on watermark performance. **Left:** Score gap heatmap across 28 configurations, gap decreases with higher overlap and lower states. **Right:** Detection rate degradation curves showing higher state counts are more sensitive to overlap.

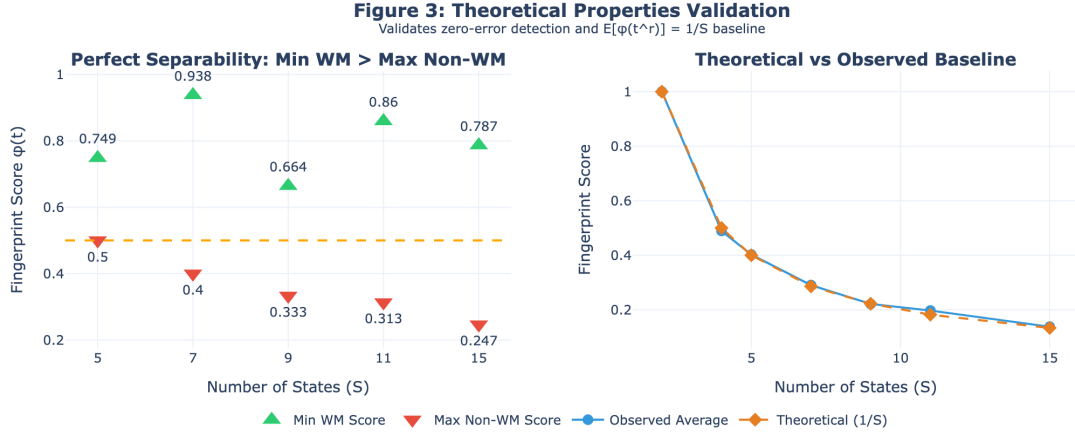


Figure 5: Theoretical properties validation. **Left:** Perfect separation for $S \geq 5$, minimum WM score exceeds maximum non-WM score. **Right:** Observed vs theoretical baseline $\mathbb{E}[\phi] = 1/S$, validating Theorem 4.1.

D. Implementation Details

Hash function: SHA-256 via Python hashlib. **State assignment:** $\text{hash}(\text{key}, \text{token_id}) \bmod S$. **Model:** Llama-3.2-3B-Instruct. **Generation:** greedy decoding ($T = 0$), max 200 tokens. **Reproducibility:** seed 42, key curated_wiki_dataset_2024.

E. Cross-Model Validation

To verify that MCL watermarking generalizes across model architectures and transition topologies, we conducted experiments on Mistral-7B-v0.3 and GPT-2-XL using identical protocols (173 Wikipedia prompts). We tested both soft cycle and clockwork transition matrices.

Table 5: Mistral-7B-v0.3 results (173 samples).

S	ρ	Topology	Avg ϕ	Detection	PPL
2	0%	soft cycle	1.000	100%	1.35
2	0%	clockwork	0.992	100%	2.05
2	5%	soft cycle	1.000	100%	1.35
2	5%	clockwork	0.876	99.4%	1.98
2	10%	soft cycle	1.000	100%	1.35
2	15%	soft cycle	1.000	100%	1.35
4	0%	soft cycle	0.992	100%	1.85
4	5%	soft cycle	0.847	98.8%	1.93
4	10%	soft cycle	0.749	97.1%	2.01
4	15%	soft cycle	0.673	91.3%	2.13
7	0%	soft cycle	0.985	100%	1.71
7	5%	soft cycle	0.607	69.9%	2.04
7	10%	soft cycle	0.499	49.7%	2.27
7	15%	clockwork	0.314	8.1%	2.38
11	0%	soft cycle	0.990	100%	2.33
11	5%	soft cycle	0.431	28.9%	2.10

Table 6: GPT-2-XL results (173 samples).

S	ρ	Topology	Avg ϕ	Detection	PPL
2	0%	soft cycle	1.000	100%	1.52
2	5%	soft cycle	1.000	100%	1.52
2	10%	soft cycle	1.000	100%	1.52
2	15%	soft cycle	1.000	100%	1.52
4	0%	soft cycle	1.000	100%	2.14

Summary. All three models (Llama-3.2-3B, Mistral-7B, GPT-2-XL) achieve 100% detection at $\rho = 0\%$ across all tested state counts and both transition topologies (soft cycle and clockwork). Mistral-7B exhibits notably lower perplexity (1.71 vs 4.20 for Llama at $S = 7$), suggesting larger vocabulary sizes reduce partitioning constraints. The overlap sensitivity pattern ($\rho \geq 5\%$ degrades detection) is consistent across architectures and topologies, confirming theoretical predictions hold regardless of model or transition matrix choice.