ChainMark: Model-Free LLM Watermarking with Closed-Form Calibration

Anonymous Authors[1]

Abstract

Regulatory regimes such as the EU AI Act mandate machine-readable marking of synthetic text, but existing watermark detectors rely on the generating LM and on heuristic thresholds with no closed-form calibration. We introduce *ChainMark*, an active watermark that partitions the vocabulary into S states via keyed SHA-256 and forces a hard Markov transition on a fraction ρ of positions; the detector replays the partition from the same key in $O(n)$ hash operations, with no LM access. We derive a closed-form $S^\star(n, \rho, \alpha)$ mapping a target FPR, text length, and budget to the minimum state count (Theorem A.2), prove a universal robustness threshold $\delta^\star = 1 - 1/\sqrt{2} \approx 29.3\%$ that is invariant in (S, ρ, n) (Theorem A.4), and generalise both to any k-regular transition topology (Theorem A.6). Across three instruction-tuned LLMs and four domains, ChainMark strictly dominates KGW and SWEET under translation and random-substitution attacks at matched budget; a one-corpus empirical recalibration restores the 1% target FPR on natural-language text (Section 6).

1. Introduction

EU AI Act Article 50 (European Parliament and Council, 2024) requires machine-readable marking of AI-generated text by August 2026, but the technical substrate is unsettled. Existing watermarks have three structural problems for regulator-facing audit. *(i) LM-bound detection.* KGW (Kirchenbauer et al., 2023) and its entropy-gated variants SWEET (Lee et al., 2024) and EWD (Lu et al., 2024) require the generating LM (or a faithful proxy) at detection time, to recompute the green-list bias or to weight tokens by entropy, so a third-party auditor cannot verify a transcript from the key alone. *(ii) No closed-form calibration.* None of these schemes inverts a regulator-side specification (target FPR α, text length $n_{\min}$, budget ρ) into a deployer-side configuration; the bias magnitude δ_{KGW} is set heuristically. *(iii) Brittle robustness.* Empirically, KGW and SWEET drop from 74.8%/71.8% TPR clean to 19.4%/18.0% after ZH back-translation at matched budget (Table 1). Distortion-free schemes (Aaronson, 2023; Kuditipudi et al., 2024) preserve quality but require LM access or the key sequence at detection; undetectability-frontier results (Christ et al., 2024) yield no deployable construction; passive detectors (Mitchell et al., 2023; Tian, 2023) expose no controllable FPR.

ChainMark. We propose *ChainMark*, an active watermark with a **model-free** detector and a **closed-form calibration formula**. A keyed SHA-256 hash partitions the vocabulary into S equivalence classes (states), and a fixed cycle over those states defines a Markov transition pattern; on a fraction ρ of generated positions, the language-model logits are masked to keep only tokens whose state is the legal successor, so the produced text walks the chain by construction. Detection re-derives every token's state from the key in $O(n)$ hash operations and counts how often consecutive tokens follow the cycle, no LM access, no proxy weights, no learned components. The detector's null law inverts in closed form to a calibration map $S^\star(n_{\min}, \rho, \alpha)$ (Theorem A.2, Section A): a regulator fixes the deployment regime and reads off the minimum admissible state count. ChainMark also enjoys a *universal* robustness floor $\delta^\star = 1 - 1/\sqrt{2} \approx 29.3\%$ (Theorem A.4), independent of (S, ρ, n), and empirically retains 72.8% TPR after ZH back-translation where KGW/SWEET drop to $\sim$19%. Figure 1 summarises the construction; the formal treatment is in Section 4 and Section 5.

Gating as a design dial. A second design choice — *which* of the n positions to watermark — is left as a parameter we call the gate. Existing schemes implicitly fix this choice: SWEET and EWD watermark only high-entropy positions, on the intuition that those positions contribute the bulk of the detection statistic. We adopt the high-entropy gate by default for parity with SWEET and compare ChainMark against KGW and SWEET at matched budget on multi-

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

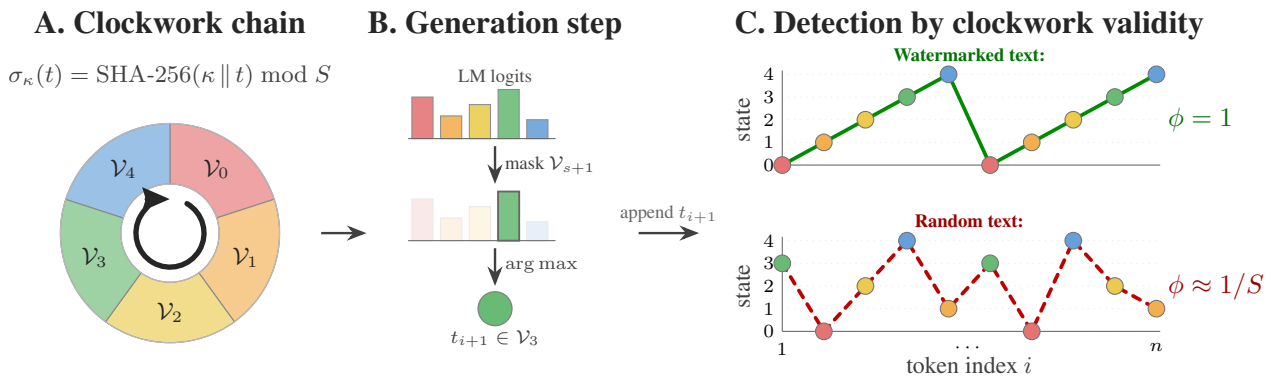


Figure 1. **ChainMark at a glance. (A)** A keyed SHA-256 hash σ_κ partitions the vocabulary into S disjoint sets $\mathcal{V}_0, \ldots, \mathcal{V}_{S-1}$ ordered into a clockwork cycle $s \to s{+}1 \bmod S$ (here $S{=}5$). **(B)** At a watermarked position with current state s, the language-model logits are masked to keep only the next partition $\mathcal{V}_{s+1}$, and the next token is the masked $\arg\max$. **(C)** Detection re-derives every token's state from the key κ in $O(n)$ hash operations: a watermarked sequence walks the chain by construction (top, validity rate $\phi = 1$), whereas unwatermarked text visits states uniformly and is valid only at rate $1/S$ in expectation (bottom).

ple instruction-tuned LLMs across four text domains under three attack conditions (*clean*, random substitution, and ZH back-translation; Section 6).

Contributions.

1. **A model-free watermarking primitive.** We introduce ChainMark, a discrete hard-constraint watermark with cryptographic state partitions, clockwork transitions, and $O(n)$ model-free detection (Section 4).
2. **Closed-form calibration.** We derive an operator-facing calibration map taking a target false-positive rate, text-length floor, and budget to the minimum admissible state count (Theorem A.2, Section A), and a universal robustness threshold that is invariant to (S, ρ, n) (Theorem A.4). Both adopt the midpoint decision convention used throughout the paper.
3. **k-regular generalisation.** Both the detection law and the robustness threshold extend verbatim to any k-regular doubly-stochastic transition topology (Theorem A.6); clockwork is the $k{=}1$ instance.
4. **Empirical head-to-head comparison.** We evaluate ChainMark against KGW and SWEET at matched budget across multiple instruction-tuned language models and four text domains under translation and random-substitution attacks (Section 6), and show that ChainMark retains substantially more detection signal than the baselines (Subsection 6.2, Subsection D.3).
5. **Empirical anchor of the calibration formula.** An independent state-count sweep recovers the closed-form $S^\star$ ordering to within one state and motivates the empirical-SD threshold recalibration that closes the residual FPR gap (Subsection 6.5, Subsection D.2).

Roadmap. Section 2 situates ChainMark within active and passive watermarking. Section 3 fixes notation; Section 4 specifies the construction; Section 5 states the calibration, robustness, and k-regular results; and Section 6 reports the empirical comparison. Proofs and additional derivations are in Section A.

2. Related Work

Passive detection. DetectGPT (Mitchell et al., 2023) exploits log-probability curvature; GPTZero (Tian, 2023) combines perplexity and burstiness. Both read statistical traces in existing text, with no controllable false-positive frontier and sharp degradation under paraphrase (Krishna et al., 2023; Sadasivan et al., 2023). They cannot serve as the technical substrate for Article 50 because neither regulator nor deployer can set the detection regime in advance.

Active watermarking. Green–red-list watermarking (Kirchenbauer et al., 2023) biases next-token logits toward a hashed green list and detects via a z-test on green-token frequency. The scheme requires temperature sampling and degrades under paraphrase, and offers no analytic inversion from a regulatory parameter back to a bias magnitude. ChainMark's fingerprint ϕ is structurally a Bernoulli$(1/S)$ z-test, so the calibration formula (8) is the analogue KGW lacks.

Distortion-free schemes. Aaronson (2023)'s Gumbel-max construction and Kuditipudi et al. (2024)'s inverse-transform sampling watermark are *distortion-free*: they preserve the LM marginal in expectation over the key. Christ et al. (2024) prove cryptographic indistinguishability under sufficient entropy. These schemes beat ChainMark on quality strictly (zero expected KL versus ChainMark's $\geq \rho \log S$). The trade is that they require either LM access or the key sequence at detection time; ChainMark is the distorting alternative that buys model-free third-party audit

and the closed-form $(\alpha, n, \rho) \mapsto S^\star$ map.

Entropy-adaptive schemes: gates vs. detectors. SWEET (Lee et al., 2024) and EWD (Lu et al., 2024) both restrict watermarking signal to *high-entropy* positions, but at different layers. SWEET applies a binary gate at generation time (restricting where the green-list bias is added); EWD weights positions by entropy at detection time, approximating the log-likelihood ratio of Theorem A.9. ChainMark is gate-agnostic in its mathematical specification: high-entropy, low-entropy, and surprisal-gap gates are all admissible. For the empirical comparison in Section 6 we adopt the high-entropy gate, matching SWEET's gating policy at the same budget; the operational dial we vary in our matched-budget head-to-head is therefore ρ, not the gate identity.

Technical AI governance. Article 50 of the EU AI Act (European Parliament and Council, 2024) and the EU's draft Code of Practice on Transparency (European Commission AI Office, 2025) require machine-readable marking of AI-generated content; the OECD Hiroshima Process (OECD, 2024) provides a parallel international framework. These instruments leave the detection regime (FPR, quality floor, robustness) implicit. ChainMark's contribution, beyond the watermarking scheme itself, is that it surfaces these parameters analytically, producing a policy-to-configuration map auditable by third parties (Section 7).

3. Preliminaries

Notation. Let $\mathcal{V}$ denote the vocabulary of an autoregressive language model, $V = |\mathcal{V}|$, and write $\mathbf{t} = (t_1, \ldots, t_n) \in \mathcal{V}^n$ for a token sequence of length n. The deployer and the verifier share a secret key $\kappa \in \{0,1\}^*$ and the same tokenizer; the verifier never sees the language model. Let $\mathcal{H} : \{0,1\}^* \to \{0,1\}^{256}$ be SHA-256, modelled as a random oracle in the security analysis (Theorem A.8). We write $[m] = \{0, \ldots, m-1\}$ and z_α for the one-sided standard-normal quantile at level α.

Threat model. The adversary observes watermarked text and may paraphrase, translate, edit, splice, or delete tokens before publication; we model these as a per-token modification rate $\delta \in [0, 1)$. The adversary is computationally bounded and has no access to κ, to the language model, or to the deployed detector as a queryable oracle at detection time (Section 7 discusses the adaptive-oracle adversary that lies outside this model). The verifier sees only token text and runs in $O(n)$ hash operations.

Detection problem and quality. Detection is a binary hypothesis test on text alone, H_0 (non-watermarked) vs. H_1 (watermarked under key κ), at a fixed false-positive rate α, since false positives are operationally the costly error. The watermark distribution must remain fluent under standard greedy or low-temperature decoding, which we control via the per-token KL divergence to the original next-token distribution and report as perplexity inflation (Section 4).

4. The ChainMark Scheme

4.1. State Partition

For each token id $v \in \mathcal{V}$ define

$$\sigma_\kappa(v) \;=\; \mathcal{H}(\kappa \,\|\, v) \bmod S, \tag{1}$$

which deterministically maps $\mathcal{V}$ onto S equivalence classes $\mathcal{V}_s = \{v : \sigma_\kappa(v) = s\}$, $s \in [S]$. Under the random-oracle model σ_κ is uniform and independent across distinct tokens, so $\mathbb{E}[|\mathcal{V}_s|] = V/S$; knowledge of σ_κ on queried tokens reveals nothing about unqueried ones (Theorem A.8).

4.2. Transition Topology

A successor function $\Sigma_k : [S] \to \binom{[S]}{k}$ assigns each state a k-element set of legal next states. We require Σ_k to be k-regular in both out- and in-degree; column-regularity is essential for the null-variance and robustness arguments (Theorem A.6, Lemma A.1). The default is the *clockwork* chain at $k = 1$, $\Sigma_1(s) = \{(s{+}1) \bmod S\}$: deterministic, periodic of period S, uniform stationary distribution, and the smallest random baseline $1/S$ among 1-regular topologies. We additionally evaluate the *soft-cycle* variant $k = 2$, $\Sigma_2(s) = \{(s{+}1) \bmod S, (s{+}2) \bmod S\}$, which doubles the per-state successor budget at the cost of raising the random baseline to $2/S$ (Section 6).

4.3. Generation

At step i with current state $s_i = \sigma_\kappa(t_i)$, ChainMark forms the legal token set $\mathcal{V}^\star_{s_i} = \bigcup_{s' \in \Sigma_k(s_i)} \mathcal{V}_{s'}$ and masks the LM logits outside $\mathcal{V}^\star_{s_i}$ to $-\infty$. At gated positions ChainMark picks the argmax over the masked logits; at ungated positions it samples from the original distribution. Each step costs one LM forward pass plus an $O(V)$ logit mask (precomputed once per state).

4.4. Watermark Budget and Entropy Gate

The mask is applied only at a subset of positions; the *watermark budget* is $\rho = \mathbb{E}_i[g_i] \in [0, 1]$. Let $H_i = -\sum_t p_i(t) \log p_i(t)$ be the next-token entropy and $\Delta_i = p_i^{(1)} - p_i^{(2)}$ the gap between the top two probabilities.

Definition 4.1 (Gate family)**.** G_{all}: $g_i = 1$; $G_{H_{\text{high}}}(\tau)$: $g_i = \mathbb{1}[H_i > \tau]$; $G_{H_{\text{low}}}(\tau)$: $g_i = \mathbb{1}[H_i < \tau]$; $G_\Delta(\tau)$: $g_i = \mathbb{1}[\Delta_i < \tau]$.

Our default is $G_{H_{\text{high}}}$, which masks only the model's uncertain positions and matches the high-entropy gating policy

Algorithm 1 ChainMark Watermark Embedding

Input: prompt p, key κ, states S, successor Σ_k, gate $g(\cdot)$, max tokens N
$\mathbf{t} \leftarrow \mathrm{Tokenize}(p)$; $s \leftarrow \sigma_\kappa(\mathbf{t}_{-1})$
for $i = 1$ **to** N **do**
  $p_i \leftarrow \mathrm{softmax}(\mathcal{M}(\mathbf{t}))$
  $\mathcal{V}^\star \leftarrow \bigcup_{s' \in \Sigma_k(s)} \mathcal{V}_{s'}$
  if $g(p_i) = 1$ **and** $p_i(\mathcal{V}^\star) > 0$ **then**
    $t_i \leftarrow \arg\max_{t \in \mathcal{V}^\star} p_i(t)$
  else
    $t_i \leftarrow \arg\max_t p_i(t)$
  end if
  $\mathbf{t} \leftarrow \mathbf{t} \,\|\, t_i$; $s \leftarrow \sigma_\kappa(t_i)$
  if $t_i = \texttt{EOS}$ **then**
    break
  end if
end for
return $\mathrm{Decode}(\mathbf{t})$

of SWEET (Lee et al., 2024) at the same budget; Section 6 reports the head-to-head. The threshold τ is calibrated by quantile-matching on a held-out pilot so the realised gate rate tracks ρ. $G_{H_{\text{low}}}$ is the *anti-SWEET* ablation, and G_Δ targets near-tied top-two positions where the greedy substitution regret is bounded pointwise by Δ_i.

4.5. Detection

The verifier tokenises the candidate text, re-derives states by re-applying σ_κ, and computes the fraction of legal transitions

$$\phi(\mathbf{t}) \;=\; \frac{1}{n-1} \sum_{i=1}^{n-1} \mathbb{1}[\sigma_\kappa(t_{i+1}) \in \Sigma_k(\sigma_\kappa(t_i))] \,. \quad (2)$$

With $p_0 = k/S$, the random-oracle null gives $\mathbb{E}[\phi] = p_0$ with pairwise-zero covariance for any k-regular Σ_k (Lemma A.1); the watermarked mean gap is $\rho(1 - p_0)$ (Theorem A.2, Theorem A.6). The verifier reports the standardised score $z = (\phi - p_0)/\sqrt{p_0(1-p_0)/(n-1)}$ and declares H_1 when $z > z_\alpha$. The closed-form calibration of Theorem A.2 (proved in Section A) selects S given target FPR α, budget ρ, and length n; the same identities yield robustness under per-token modification rate δ (Theorem A.4).

Algorithm 2 runs in $O(n)$ hash operations without LM access, making ChainMark deployable as a third-party audit primitive.

5. Theoretical Properties

This section gives an intuitive tour of three core results (detection, robustness, k-regular generalisation) plus two supporting properties (security against an oracle-blind adversary, and an LM-aware locally most powerful detector). Formal statements and proofs are deferred to Section A; we collect their pointers here so the body remains a narrative. Empirical anchoring of Theorem A.2 appears in Subsection 6.5.

Algorithm 2 ChainMark Watermark Detection (model-free)

Input: text x, key κ, states S, successor Σ_k, level α
$\mathbf{t} \leftarrow \mathrm{Tokenize}(x)$; $c \leftarrow 0$
for $i = 1$ **to** $n - 1$ **do**
  if $\sigma_\kappa(t_{i+1}) \in \Sigma_k(\sigma_\kappa(t_i))$ **then**
    $c \leftarrow c + 1$
  end if
end for
$\phi \leftarrow c/(n-1)$; $p_0 \leftarrow k/S$
$z \leftarrow (\phi - p_0)/\sqrt{p_0(1-p_0)/(n-1)}$
return $(z > z_\alpha,\ \phi,\ z)$

Detection bound and calibration. Under the null hypothesis that the input is a uniformly random token sequence, the fingerprint score ϕ of Equation 2 is approximately Gaussian with mean $1/S$ and variance $(1/S)(1-1/S)/(n-1)$, while a watermarked sequence with gate density ρ shifts the mean to $1/S + \rho(S-1)/S$. Standardising gives the closed-form z-statistic $z(\rho, S, n) = \rho\sqrt{(S-1)(n-1)}$. Inverting at false-positive level α under the midpoint threshold convention yields a state-count rule that practitioners can read off directly,

$$S^\star(n, \rho, \alpha) \;=\; \left\lceil \frac{4\, z_\alpha^2}{\rho^2 (n-1)} + 1 \right\rceil .$$

The factor-of-4 comes from evaluating the standardised signal at the midpoint between the null and alternative means. This gives the regulator a $2\times$ safety margin relative to the one-sided z_α test reported in our empirical tables; those tables run at a tighter operating threshold than $S^\star$ requires, so empirical TPR exceeds the conservative midpoint prediction. Formal statement and proof in Theorem A.2 of Section A; the corresponding lookup table is Table 4.

Empirical-SD calibration recipe. The closed-form z_α assumes i.i.d. uniform tokens over the SHA-256 partition; on natural-language text this drifts (empirically 1.7–2.0% at the nominal 1% target, Table 3). The fix is a one-corpus recalibration: estimate the empirical mean and SD of z on a non-watermarked sample of the deployed LM (or a domain-matched corpus) and use $z^\star = \hat{\mu} + z_\alpha\, \hat{\sigma}$ as the operating threshold. Under the same Gaussian-tail null this restores the target FPR (1.17% at $\alpha = 1\%$ on our $n = 3000$ pooled corpus) without changing $S^\star$ or touching the watermarked side, so TPR is unaffected. The empirical z-null is mildly leptokurtic, so for tighter α or large S the

empirical-quantile drop-in $z^\star = \hat{F}^{-1}(1-\alpha)$ is the robust alternative. Formal recipe in Proposition A.3; empirical evaluation in Subsection D.2.

Universal robustness threshold. Suppose an adversary independently replaces each token with probability δ by a fresh token whose state is uniform on $[S]$. Both the watermarked-pair signal and the midpoint-threshold gap scale by the same affine factor $\rho(S-1)/S$, so the post-attack-to-pre-attack z ratio collapses to $(1-\delta)^2$, independently of (S, ρ, n). The critical edit fraction at which detection fails is therefore the universal constant $\delta^\star = 1 - 1/\sqrt{2} \approx 0.293$. Formal statement and proof in Theorem A.4.

k-regular generalisation. The clockwork transition is the simplest member of a broader family: any k-regular adjacency T (every state has exactly k allowed successors and predecessors) yields a valid ChainMark scheme. Replacing the random baseline $1/S$ by $p_0 = k/S$ gives the same calibration identities, the same midpoint critical fraction $\delta^\star$, and a single quality-versus-detection dial in k. Column-regularity is load-bearing: it is what makes adjacent indicator pairs have zero covariance under the null, so the variance formula carries through. Formal statement and proof in Theorem A.6. The random baseline $p_0 = k/S$ relies on the SHA-256 partition being approximately uniform across the vocabulary, an assumption we inherit from the standard avalanche property of cryptographic hashes.

Security and an LM-aware optimal detector. The state map $\sigma_\kappa(t) = \mathcal{H}(\kappa \,\|\, t) \bmod S$ is modelled as a random oracle: an adversary without access to κ, the LM, or reference text cannot predict $\sigma_\kappa(t^\star)$ on a fresh token with non-negligible advantage in the key min-entropy, and therefore cannot statistically distinguish ChainMark output from random text (Theorem A.8; computational, not information-theoretic, and excluding adversaries with LM access). A verifier that *does* have LM access can sharpen detection by replacing ϕ with an entropy-weighted log-likelihood ratio that is asymptotically locally most powerful under the random-oracle product-form approximation (Theorem A.9). Detector pseudocode appears as Algorithm 3.

6. Experiments

We run four experiments on three instruction-tuned 7–8B LLMs across four domains. Per-prompt records stream to JSONL for reproducibility. Figure 2 traces a single ChainMark generation end-to-end on a wiki-domain prompt: the model receives the user prompt, ChainMark masks logits at gated positions to keep only states $(s+1) \bmod S$, and the detector re-derives every token's state from the key alone and reports a z-score.

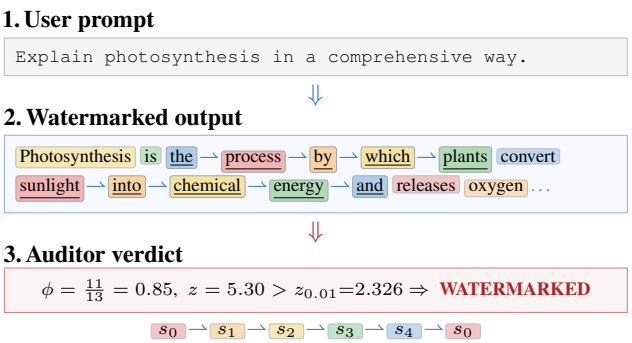


Figure 2. **End-to-end ChainMark trace** on a wiki prompt (Llama-3.1-8B-Instruct, $S{=}5$, $\rho{=}0.5$, $\sim$16-token excerpt). Every output tile inherits its colour from $\sigma_\kappa(t) \in \{0,1,2,3,4\}$ (legend at bottom). Underlined tokens are gated steps where ChainMark masked logits to enforce $s_{i+1}{=}(s_i{+}1) \bmod 5$ and the → arrows trace the forced clockwork walk; un-underlined tokens are free-sampled at $T{=}0.7$. The auditor in panel 3 re-derives the entire colour sequence from κ alone, counts $11/13$ valid transitions (green bar), and emits (ϕ, z) in $O(n)$ hash operations, no LM forward pass.

6.1. Setup

Models. Three instruction-tuned, openly licensed checkpoints in the 7–8B parameter range: Llama-3.1-8B-Instruct (Grattafiori et al., 2024), Qwen-2.5-7B-Instruct (Yang et al., 2024), and Mistral-7B-Instruct-v0.3 (Jiang et al., 2023). The same secret key κ is fixed across every run.

Domains. Four prompt domains stress different entropy regimes: *code* (HumanEval problem stems), *factual* (short closed-form knowledge prompts), *wiki* (open-ended "Explain X..." prompts over a curated concept list), and *writing* (creative-completion prompts).

Generation and detection. Temperature $T = 0.7$, top-$p = 1$, token budget $n = 200$ at every cell, with one deterministic seed so the same indices are watermarked, attacked, and detected across methods. Detection uses the threshold z_α from Theorem A.2, reused across methods for cross-method consistency; the empirically calibrated 1%-FPR head-to-head (Subsection D.3) confirms the resulting TPR ranking is not an FPR artefact.

ChainMark configuration. Unless stated otherwise, ChainMark runs use $S = 5$ states, target gate budget $\rho = 0.5$, the high-entropy gate $G_{H_{\text{high}}}$, and clockwork ($k{=}1$) topology.

Baselines. *KGW* (Kirchenbauer et al., 2023) at $\gamma = 0.5$, logit bias $\delta_{\text{KGW}} = 2$ (the canonical operating point reported in the original paper at this γ). *SWEET* (Lee et al., 2024) is reproduced as a matched-budget re-implementation: same $\gamma = 0.5$ and $\delta_{\text{KGW}} = 2$ but the green-list bias is applied only at the top-$\rho{=}0.5$ fraction of positions by token-level entropy, so the budget aligns with ChainMark's ρ. Per-domain prompt-and-generation examples (with a

Table 1. **Headline aggregate** at the analytic threshold $z_\alpha = 2.326$ ($\alpha = 0.01$, Gaussian-tail null), aggregated over 3 LLMs $\times$ 4 domains $\times$ 100 prompts at n=200 tokens per cell. PPL is the median across the 12 cell medians on the watermarked model's own logits (caveats in Section 7). Empirical FPR drift at this threshold appears separately in Table 3. Bold = best per column.

Method	$\text{TPR}_{\text{clean}}$ (%)	TPR_{rnd} (%)	TPR_{ZH} (%)	PPL↓
ChainMark	**100.0**	**99.9**	**72.8**	3.66
KGW	74.8	55.1	19.4	1.93
SWEET	71.8	51.2	18.0	**1.80**

token-level gating walkthrough) appear in Section F; the end-to-end trace in Figure 2 above gives one illustration on the wiki prompt `Explain photosynthesis in a comprehensive way..`

6.2. Matched-Budget Head-to-Head

We compare ChainMark against KGW and SWEET on 1 200 generations per method (3 models, 4 domains, 100 prompts each), with 200 tokens per generation, and score each generation under three conditions: clean, random-substitution at δ_{eff}=0.20, and NLLB-200 EN→ZH→EN back-translation. Aggregated across all 1 200 cells per method (Table 1, Figure 3a), ChainMark detects 100% of clean and 99.9% of randomly perturbed generations and retains 72.8% TPR after ZH back-translation, where KGW and SWEET drop from ∼75%/∼72% clean to 19.4%/18.0% post-attack. The gap is consistent across every (model, domain) cell (Figure 4); ChainMark matches or leads every clean cell and strictly leads every post-ZH cell. ChainMark pays a $\approx 2\times$ self-perplexity gap (3.66 vs 1.80/1.93) for this gain. SWEET is an entropy-gating ablation of KGW with the gating policy held fixed against ChainMark.

6.3. Translation Robustness

Each ChainMark-watermarked output (3 models $\times$ 4 domains $\times$ 25 prompts $\times$ n=200 tokens, 1 200 generations total) is round-tripped through NLLB-200 (NLLB Team, 2022) on four pivot languages (FR, DE, RU, ZH), and we compute the empirical δ_{eff} per output via token-edit distance. Plotting the post-attack-to-clean z ratio against δ_{eff} (Figure 3b) shows the observed ratios sit at or above the closed-form $(1-\delta_{\text{eff}})^2$ diagonal across pivots, models, and domains, which empirically anchors the universal half-life $\delta^\star = 1 - 1/\sqrt{2}$ from Theorem A.4 and confirms the bound is conservative rather than tight. The claim is stronger than synthetic random-substitution can support, since NLLB-200 produces semantically coherent rewrites rather than i.i.d. token noise.

Table 2. **k-regular ChainMark at k=2 (soft-cycle), S=5, ρ=0.5.** Predicted random baseline $p_0 = 0.40$; wiki domain, 100 prompts. The clean $\bar{\phi}_{\text{cl}}$ values (≥ 0.88) sit well above p_0 (Theorem A.6); the post-attack z ratio re-certifies Theorem A.4 at k=2 (Section A).

Model	$\bar{\phi}_{\text{cl}}$	$\bar{z}_{\text{cl}}$	TPR_{cl}	$\bar{z}_{\text{ZH}}$	TPR_{ZH}
Llama-3.1-8B	0.883	13.91	100%	3.19	54%
Mistral-7B	0.965	16.23	100%	3.91	70%
Qwen-2.5-7B	0.934	15.38	100%	4.12	71%

Table 3. **Empirical anchor of the calibration map $S^\star(n, \rho, \alpha)$ from Theorem A.2.** For each S, aggregated across 3 LLMs $\times$ 4 domains $\times$ 25 prompts ($\approx$ 300 watermarked + 300 non-watermarked). $z_{\text{wm}}^{\text{pred}} = \rho\sqrt{(S-1)(n-1)}$. Watermarked z overshoots theory by 50–80%; non-watermarked FPR_{nwm} exceeds the α=1% target at large S due to non-i.i.d. structure in natural-language token sequences. The last two columns apply the empirical-quantile recalibration of Proposition A.3 per-S on the same NWM corpus: $\text{FPR}_{\text{nwm}}^{\text{recal.}}$ lands uniformly below α=1%, while $\text{TPR}_{\text{wm}}^{\text{recal.}}$ stays at $\geq 95\%$ on the watermarked side (the recalibration only lifts the threshold, so detection power is essentially preserved).

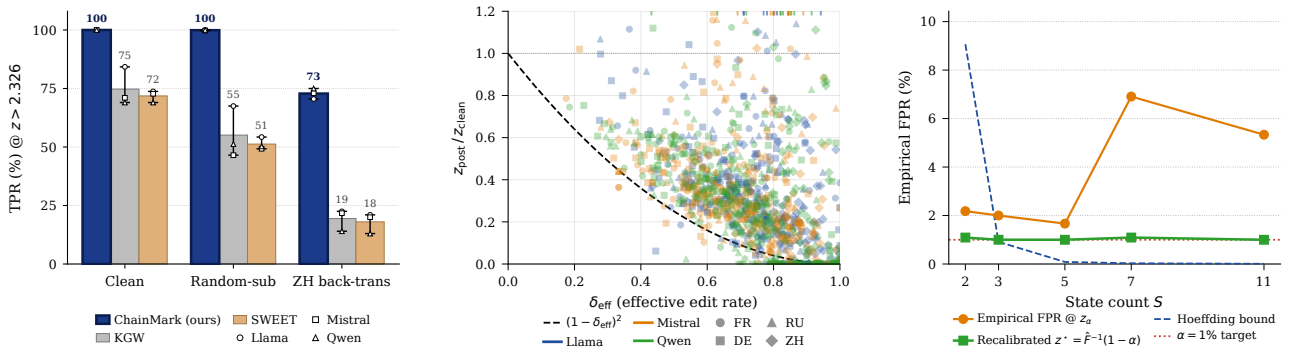
S	$\bar{z}_{\text{wm}}^{\text{obs}}$	$z_{\text{wm}}^{\text{pred}}$	TPR_{wm}	FPR_{nwm}	$\text{TPR}_{\text{wm}}^{\text{recal.}}$	$\text{FPR}_{\text{nwm}}^{\text{recal.}}$
2	10.87	7.05	100%	2.00%	98.7%	0.67%
3	17.41	9.97	100%	2.00%	100.0%	0.67%
5	25.38	14.11	100%	1.67%	100.0%	0.34%

6.4. k-Regular at k=2

Instantiating $k = 2$ (each state has two allowed successors) at $S = 5$ predicts $p_0 = k/S = 0.40$. We run 100 wiki-domain prompts on each of the three models with ZH back-translation as the only attack. Table 2 reports clean $\bar{\phi}_{\text{cl}} \geq 0.88$ (well above the $p_0 = 0.40$ prediction) and a post-attack z-score that sits above the $(1-\delta_{\text{eff}})^2$ floor of Theorem A.4 (observed ratio 0.23–0.27 vs. predicted 0.07–0.09, i.e. the bound is conservative as in Subsection 6.3), jointly re-certifying Theorem A.6 and Theorem A.4 at a non-trivial k.

6.5. Calibration Anchor

We sweep $S \in \{2, 3, 5\}$ across the three models with 4 domains $\times$ 25 prompts at n=200, recording both a watermarked generation (ChainMark at the canonical operating point, varying only S) and a non-watermarked baseline drawn from the same prompt. Non-watermarked outputs feed the empirical FPR estimate at the analytic z_α; watermarked outputs feed the TPR. Table 3 shows watermarked z overshoots the closed-form prediction $\rho\sqrt{(S-1)(n-1)}$ by 50–80%, so TPR is at 100% throughout, while empirical FPR sits above the $\alpha = 1\%$ target by $\sim$1 pp (Figure 3c, 1.7–2.0%). The closed form gives the right $S^\star$ ordering to within one state; for strict 1% FPR a deployer applies the per-S empirical-quantile threshold recalibration of Proposition A.3 (Section 7, Subsection D.2).



(a) **Head-to-head TPR.** ChainMark vs KGW vs SWEET, aggregated over 3 LLMs × 4 domains.

(b) **Translation robustness.** Empirical decay ratio under NLLB-200 EN→{FR,DE,RU,ZH}→EN with the closed-form $(1-\delta_{\text{eff}})^2$ diagonal (Theorem A.4); n=100 per (model, pivot).

(c) **Calibration anchor.** Empirical FPR vs state count S, with the closed-form Hoeffding upper bound (Theorem A.2).

Figure 3. **Experimental headlines. (a)** head-to-head TPR (§6.2); **(b)** translation-decay curve (§6.3); **(c)** FPR calibration anchor (§6.5).

7. Discussion

7.1. When to Use What

The calibration $S^\star(n_{\min}, \rho, \alpha)$ of Theorem A.2 reduces operational deployment to three inputs: a minimum text length $n_{\min}$, a target false-positive rate α, and a watermark budget ρ. For long-form content the calibration anchors comfortably at small S with ρ near $1/2$; the head-to-head in Subsection 6.2 shows ChainMark substantially exceeds KGW and SWEET in detection power at the analytical $z_\alpha = 2.326$ threshold (the apples-to-apples empirical-FPR$= 1\%$ recalibration in Subsection D.3 preserves the lead vs. KGW; the SWEET row is deferred since the SWEET null was not collected on the FPR corpus), while incurring a $\approx 2\times$ higher self-perplexity (3.66 vs 1.80–1.93; Table 1). For short outputs $S^\star$ rises sharply with stricter α, so deployers should either tolerate a higher state count or shift to a longer floor. When cross-lingual rewrite or random substitution dominates the expected attack profile (Subsection 6.2, Subsection 6.3), the universal robustness threshold $\delta^\star$ of Theorem A.4 bounds the worst case across all gating choices.

7.2. Empirical SD Recalibration Closes the FPR Gap

Recomputing the detection threshold from the empirical SD of z on a non-watermarked corpus gives $z^\star = \hat{\mu} + z_\alpha\,\hat{\sigma}$, which brings empirical FPR from 2.1% to 1.2% at $\alpha = 1\%$ while preserving TPR$= 100\%$ on the pooled n=3000 FPR corpus (Subsection D.2); the per-S empirical-quantile drop-in in Table 3 costs at most 1.3 pp of TPR ($\text{TPR}_{\text{wm}}^{\text{recal.}} \geq 98.7\%$ across $S \in \{2, 3, 5\}$). The plug-in is calibrated at $\alpha = 1\%$ but the empirical z-null is leptokurtic (excess kurtosis ≈ 1.4); for tighter $\alpha \leq 0.5\%$ the empirical-quantile recipe $z^\star = \hat{F}^{-1}(1-\alpha)$ is the robust drop-in (and is what we use to draw the recalibrated curve in Figure 3c). The apples-to-apples 1%-FPR head-to-head appears in Subsection D.3; ChainMark also admits an empirical FPR$= 0\%$ regime via threshold lifting (Subsection D.4), which KGW and SWEET cannot achieve since their watermarked and non-watermarked z distributions overlap. Failure modes we tried but did not adopt (Newey–West HAC, k-skip, stopword filtering, HDD-lite) are tabulated in Subsection D.2.

7.3. Limitations

Experimental scope. The headline grid fixes $\rho = 0.5$, KGW/SWEET at γ=0.5, δ_{KGW}=2, and decoding at T=0.7, top-p=1, n=200 with one deterministic prompt-order seed (no replicate-seed CIs). The fleet covers three open-weight instruction-tuned 7–8B models; base, smaller, and $\geq$ 10B models are deferred. We validate k-regular ChainMark at $k \in \{1, 2\}$; $k \geq 3$ is deferred. The robustness suite is translation-only: we do *not* run grammar-preserving paraphrase (e.g. DIPPER), so the strongest robustness claim is against translation and uniform substitution, not against adversarial paraphrasing. The translation evaluation now covers all 12 (model, domain) cells (Table 5), but the SWEET row of the empirically recalibrated head-to-head (Table 7) is still omitted since the SWEET null z-distribution was not collected on the FPR corpus.

Quality and detector caveats. Generation quality is self-perplexity under each model's own logits; we do not report external-LM PPL, MAUVE, or human-rater scores. The detector is "model-free" in the sense of needing no LM forward pass at detection, but it does require the same tokenizer used at generation; an auditor handed plain text alone would have to enumerate candidate tokenizers. Both Theorem A.8

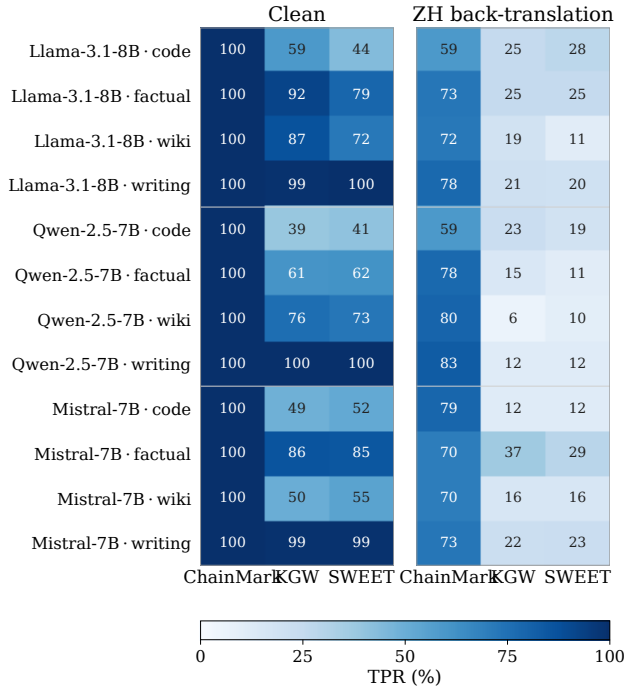


Figure 4. **Per-(model, domain) head-to-head TPR (%)** at $z_\alpha = 2.326$ (100 prompts per cell, n=200 tokens). Each row is one (model, domain) cell; left panel is clean, right panel is ZH back-translation. ChainMark matches or leads every clean cell and strictly leads every cell post-ZH-back-translation. Cells with TPR<50% are annotated with their value in white inside the cell.

and Theorem A.4 cover an oracle-blind adversary only; an attacker with adaptive detector-query access (or with σ_κ access) can choose substitution targets to hit valid transitions and is outside our threat model. Query complexity for such an adversary is open.

FPR drift and theorem scope. The empirical FPR at the analytic Gaussian-tail threshold $z_\alpha = 2.326$ drifts to 1.7–2.0% on non-watermarked LLM output (Table 3); natural-language token sequences are not i.i.d. uniform over the SHA-256 partition, so the closed-form Hoeffding envelope (Figure 3c) sits below the empirical curve. Headline TPR numbers in Table 1–Figure 4 are at the analytic threshold (cross-method consistency), not at an empirically calibrated 1% FPR per cell. Theorem A.4 itself is derived under i.i.d. uniform substitution at rate δ; NLLB back-translation (and the deferred paraphrase extension) preserves grammatical structure and correlates substitutions across positions. Our empirical curves (Figure 3b) sit *above* the $(1-\delta_{\text{eff}})^2$ slope, so the bound is conservative rather than tight, and δ_{eff} is an effective-edit-rate proxy.

8. Conclusion

We presented ChainMark, an active watermark with two operator-facing properties. Its closed-form calibration $S^\star(n_{\min}, \rho, \alpha)$ maps a regulator-facing specification (target false-positive rate, text length, budget) directly to the minimum state count, and its detector requires only the secret key, no language-model access. The scheme inherits a universal robustness floor $\delta^\star = 1 - 1/\sqrt{2} \approx 29.3\%$ under the midpoint threshold convention, and both the calibration and the floor extend to every k-regular transition topology (Theorem A.2, Theorem A.4, Theorem A.6). Head-to-head against KGW and SWEET on three instruction-tuned LLMs across four text domains at matched budget, ChainMark retains substantially more detection signal under translation and random-substitution attacks. The empirical post-attack z ratio sits above the worst-case $(1-\delta_{\text{eff}})^2$ scaling, so we read our robustness bound as conservative rather than tight (Section 6). Four directions are immediate. First, tighten the *analytic* FPR bound under non-i.i.d. natural-language statistics: the empirical-SD recalibration of Subsection D.2 closes the deployment gap, but the closed form remains loose at large S (Table 3). Second, extend the protocol to longer text regimes where small $S^\star$ becomes the operative regime. Third, validate k-regular topologies at $k \geq 3$. Fourth, characterise behaviour on base (non-instruction-tuned) LLMs whose entropy profile differs from the instruct-tuned fleet studied here.

Impact Statement

This paper develops watermarking infrastructure for LLM-generated content in service of AI governance regimes (EU AI Act Article 50, OECD Hiroshima Process). The contribution is dual-use, and we flag four asymmetric harms.

Authorship tracking and chilling effects. Stronger watermarking exposes writers who relied on undetectable LLM use. Marginalised users (ESL writers, students under inequitable AI policies, whistleblowers reformulating sensitive content) bear asymmetric harm relative to incumbent users.

End-user perplexity cost. ChainMark imposes a $\approx 2\times$ self-perplexity cost (3.66 vs the 1.80–1.93 baseline range; Table 1). Deployers adopting ChainMark to satisfy a regulatory mark therefore impose a measurable quality cost on *every* user whose generation is gated, not just on adversaries who would try to evade detection.

False positives. Mislabelling human-written text as machine-generated has real reputational and legal cost. We instrument the detector with a closed-form, regulator-facing target false-positive rate α rather than an unspecified threshold; the discussed empirical-quantile recalibration (Section 7, Subsection D.2) is the recipe a deployer should run

before invoking detection on any production content.

Adaptive adversaries are out of scope. Our security and robustness theorems cover an oracle-blind adversary; a regulator deploying ChainMark must not assume the watermark survives an attacker with detector-query access or with σ_κ access (Section 7). The audit primitive is honest about non-adversarial mislabelling, not about adversarial spoofing or scrubbing under realistic API-query budgets.

Any deployment should publish the detection regime (α, $n_{\min}$, ρ, the recalibration recipe used) so users understand the conditions under which their outputs are, and are not, marked.

References

Aaronson, S. Watermarking GPT outputs. Blog post, https://scottaaronson.blog/?p=6823, 2023. Accessed January 2026.

Christ, M., Gunn, S., and Zamir, O. Undetectable watermarks for language models. In *Conference on Learning Theory (COLT)*, pp. 1042–1100, 2024.

European Commission AI Office. First draft code of practice on transparency of ai-generated content, 2025. URL https://digital-strategy.ec.europa.eu/en/library/first-draft-code-practice-transparency-ai-generated-content. Supporting implementation of EU AI Act Article 50.

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.

Grattafiori, A. et al. The Llama 3 Herd of Models, 2024. Model card: https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct.

Jiang, A. Q. et al. Mistral 7B, 2023. Model card: https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.3.

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

Krishna, K., Song, Y., Karpinska, M., Wieting, J., and Iyyer, M. Paraphrasing evades detectors of AI-generated text, but retrieval is an effective defense. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 13287–13300, 2023.

Kuditipudi, R., Thickstun, J., Hashimoto, T., and Liang, P. Robust distortion-free watermarks for language models. *Transactions on Machine Learning Research*, 2024.

Lee, T., Hong, S., Ahn, J., Hong, I., Lee, H., Yun, S., Shin, J., and Kim, G. Who wrote this code? watermarking for code generation (SWEET). In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2024.

Lu, Y. et al. EWD: Entropy-weighted watermark detection for language models. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2024.

Mitchell, E., Lee, Y., Khazatsky, A., Manning, C. D., and Finn, C. DetectGPT: Zero-shot machine-generated text detection using probability curvature. In *International Conference on Machine Learning (ICML)*, pp. 24950–24962, 2023.

NLLB Team. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*, 2022.

OECD. Hiroshima AI process: International code of conduct for advanced AI systems, 2024. URL https://www.oecd.org/digital/hiroshima-ai-process. Organization for Economic Co-operation and Development.

Sadasivan, V. S., Kumar, A., Balasubramanian, S., Wang, W., and Feizi, S. Can AI-generated text be reliably detected? *arXiv preprint arXiv:2303.11156*, 2023.

Tian, E. GPTZero: Towards detection of AI-generated text using zero-shot and supervised methods. *arXiv preprint arXiv:2301.11305*, 2023.

Yang, A. et al. Qwen2.5 Technical Report, 2024. Model card: https://huggingface.co/Qwen/Qwen2.5-7B-Instruct.

A. Proofs and Formal Statements

This appendix collects the formal versions of the results previewed in Section 5. Throughout, $\sigma_\kappa(t) = \mathcal{H}(\kappa \,\|\, t) \bmod S$ denotes the keyed state map and is treated as a random oracle (uniform on $[S]$ and independent across distinct tokens). ϕ is the fingerprint score of Equation 2, and we write $X_i = \mathbf{T}_{\sigma_\kappa(t_i),\sigma_\kappa(t_{i+1})}$ for the validity indicator at step i, with $T = \mathbf{T}^{\mathrm{clk}}$ in the clockwork case.

A.1. Null Variance Lemma

The null variance underpins both the detection z-score and the universal robustness threshold; we isolate it as a lemma.

Lemma A.1 (Null variance with zero pairwise covariance)**.** *Let $T \in \{0,1\}^{S\times S}$ be k-regular: every row and every column has exactly k ones, with $1 \le k \le S-1$, and*

set $p_0 = k/S$. Under the random oracle null (i.i.d. uniform tokens), the indicators $X_i = T_{\sigma_\kappa(t_i),\sigma_\kappa(t_{i+1})}$ satisfy $\mathbb{E}[X_i] = p_0$ and $\mathrm{Cov}(X_i, X_j) = 0$ for all $i \neq j$, hence

$$\mathrm{Var}[\phi(\mathbf{t}^r)] \;=\; \frac{p_0(1-p_0)}{n-1}. \tag{3}$$

Proof. Write $N^+(s) = \{s' : T(s,s') = 1\}$ (equivalently $\Sigma_k(s)$ from Section 4) and $N^-(s') = \{s : T(s,s') = 1\}$; by k-regularity, $|N^+(s)| = |N^-(s')| = k$ for every s, s'. Throughout this proof we operate under the random-oracle null with i.i.d. uniform tokens on $\mathcal{V}$; the empirical FPR drift observed on natural-language text (Table 3, Section 7) is the practical price of this idealisation.

Marginal mean. Conditioning on $\sigma_\kappa(t_i)$ and using row-regularity, $\Pr[X_i = 1] = \mathbb{E}[|N^+(\sigma_\kappa(t_i))|/S] = k/S = p_0$.

Non-adjacent pairs. For $|i-j| \geq 2$, X_i and X_j depend on disjoint token pairs and thus on independent state evaluations of the random oracle (assuming the underlying tokens are distinct; coincidences contribute $O(1/|\mathcal{V}|)$ and are absorbed into the random-oracle approximation). Hence $\mathrm{Cov}(X_i, X_j) = 0$.

Adjacent pairs. X_i and X_{i+1} share the token t_{i+1}. Conditioning on $\sigma_\kappa(t_{i+1}) = s'$, we have

$$\Pr[X_i = 1 \mid \sigma_\kappa(t_{i+1}) = s'] \;=\; |N^-(s')|/S \;=\; p_0,$$

$$\Pr[X_{i+1} = 1 \mid \sigma_\kappa(t_{i+1}) = s'] \;=\; |N^+(s')|/S \;=\; p_0,$$

using *column*-regularity for the first identity. Conditional on $\sigma_\kappa(t_{i+1})$, the indicators X_i and X_{i+1} depend on the disjoint random-oracle evaluations $\sigma_\kappa(t_i)$ and $\sigma_\kappa(t_{i+2})$, so they are conditionally independent whenever $t_i \neq t_{i+2}$ (the coincidence event has probability $1/|\mathcal{V}|$ under iid uniform tokens and is absorbed into the random-oracle approximation, parallel to the non-adjacent case above). Therefore

$$\Pr[X_i = X_{i+1} = 1] \;=\; \mathbb{E}_{s'}[p_0 \cdot p_0] \;=\; p_0^2,$$

and $\mathrm{Cov}(X_i, X_{i+1}) = 0$.

Variance. $\phi = (n-1)^{-1}\sum_{i=1}^{n-1} X_i$ is the average of $n-1$ Bernoulli(p_0) random variables with all pairwise covariances vanishing, giving (3). □

A.2. Detection Bound and Calibration

Theorem A.2 (Detection bound and calibration)**.** *Let $\mathbf{t}^w$ be a ChainMark-watermarked sequence of length n with watermark fraction ρ and clockwork transition over S states; let $\mathbf{t}^r$ be an i.i.d. random sequence over $\mathcal{V}$. Then*

$$\mathbb{E}[\phi(\mathbf{t}^w)] = \tfrac{1}{S} + \rho\,\tfrac{S-1}{S}, \tag{4}$$

$$\mathbb{E}[\phi(\mathbf{t}^r)] = \tfrac{1}{S}, \tag{5}$$

$$\mathrm{Var}[\phi(\mathbf{t}^r)] = \tfrac{(1/S)(1-1/S)}{n-1}, \tag{6}$$

$$z(\rho, S, n) = \rho\sqrt{(S-1)(n-1)}. \tag{7}$$

Under the midpoint threshold $\tau_{\mathrm{mid}} = \frac{1}{S} + \frac{\rho}{2}\frac{S-1}{S}$, splitting the indicators into disjoint odd/even subsequences and applying Hoeffding gives

$$\mathrm{FPR} \;\leq\; 2\exp\!\big(-2\lfloor (n-1)/2\rfloor((\rho/2)(S-1)/S)^2\big).$$

Using the Gaussian convention at the midpoint threshold, the minimum state count guaranteeing $\mathrm{FPR} \leq \alpha$ is

$$S^\star(n, \rho, \alpha) \;=\; \left\lceil \frac{4z_\alpha^2}{\rho^2(n-1)} + 1 \right\rceil. \tag{8}$$

$S^\star$ is a provisioning *bound: it is derived under the midpoint detector τ_{mid} to give the regulator a $2\times$ safety margin in the standardised mean gap. The* operating *detector deployed in Section 6 uses the tighter one-sided test $\{z > z_\alpha\}$, so empirical TPR at the operating threshold exceeds the conservative midpoint prediction whenever $S \geq S^\star(n, \rho, \alpha)$.*

Proof. Null mean. Under the random oracle, σ_κ is uniform and independent across distinct tokens, so $X_i \sim \mathrm{Bernoulli}(1/S)$ marginally and $\mathbb{E}[\phi(\mathbf{t}^r)] = 1/S$.

Watermark mean. At a gated position, Algorithm 1 sets $\sigma_\kappa(t_{i+1}) = (\sigma_\kappa(t_i) + 1) \bmod S$, so $X_i = 1$ deterministically. At an ungated position, $X_i \sim \mathrm{Bernoulli}(1/S)$. With gate density ρ,

$$\mathbb{E}[\phi(\mathbf{t}^w)] = \rho \cdot 1 + (1-\rho)\cdot\tfrac{1}{S} = \tfrac{1}{S} + \rho\tfrac{S-1}{S}.$$

Null variance. Clockwork is the $k = 1$ instance of Lemma A.1 (column-regularity is trivial since $|N^-(s')| = 1$), giving $\mathrm{Var}[\phi(\mathbf{t}^r)] = (1/S)(1-1/S)/(n-1)$.

z-score. The signed mean gap is $\rho(S-1)/S$. Dividing by the null SD,

$$\begin{aligned} z(\rho, S, n) &= \frac{\rho(S-1)/S}{\sqrt{(1/S)(1-1/S)/(n-1)}} \\ &= \rho\sqrt{(S-1)(n-1)}. \end{aligned}$$

FPR via Hoeffding. Adjacent indicators share a token and are not jointly i.i.d., so we split into $Z_1, Z_3, Z_5, \ldots$ and $Z_2, Z_4, Z_6, \ldots$, two disjoint subsequences of $m = \lfloor (n-1)/2 \rfloor$ terms, each i.i.d. Bernoulli$(1/S)$. Since ϕ is the average of these two subsequence means, $\phi \geq \tau$ implies that at least one subsequence mean is $\geq \tau$, so

by Hoeffding plus a union bound, for any $\tau > 1/S$, $\Pr[\phi \geq \tau] \leq 2\exp(-2m(\tau - 1/S)^2)$.

Calibration. By the CLT for 1-dependent sequences (the $\{X_i\}$ are 1-dependent and bounded with vanishing pairwise covariance), $\sqrt{n-1}\,(\phi - 1/S)$ is asymptotically $\mathcal{N}(0, p_0(1-p_0))$ under the null, with $p_0 = 1/S$. At the midpoint threshold $\tau_{\text{mid}} = 1/S + (\rho/2)(S-1)/S$, the null z-margin is $(\rho/2)\sqrt{(S-1)(n-1)}$. Requiring it to exceed z_α,

$$\tfrac{\rho}{2}\sqrt{(S-1)(n-1)} \;\geq\; z_\alpha \quad\Longleftrightarrow\quad S \;\geq\; \frac{4z_\alpha^2}{\rho^2(n-1)} + 1,$$

and taking the ceiling yields (8). □

Proposition A.3 (Empirical-SD recalibration recipe)**.** *Let $\mathbf{t}^{(1)}, \ldots, \mathbf{t}^{(M)}$ be M i.i.d. non-watermarked sequences from the deployed LM, and let $z_1, \ldots, z_M$ be their detection z-statistics under ChainMark with (S, ρ) fixed. Define*

$$\hat{\mu} = \tfrac{1}{M}\sum_{i=1}^{M} z_i,$$
$$\hat{\sigma}^2 = \tfrac{1}{M-1}\sum_{i=1}^{M}(z_i - \hat{\mu})^2,$$
$$z^\star = \hat{\mu} + z_\alpha\,\hat{\sigma}.$$

If the z_i are approximately Gaussian with mean $\hat{\mu}$ and SD $\hat{\sigma}$ (a calibration-side assumption that holds far better than the i.i.d. random-oracle null assumed by Theorem A.2), then declaring H_1 when $z > z^\star$ has FPR $\to \alpha$ as $M \to \infty$. When the z-null is measurably non-Gaussian (e.g. heavy-tailed at large S), the Gaussian plug-in $z^\star$ is replaced by the empirical-quantile *drop-in $z^\star_{\text{emp}} = \hat{F}^{-1}(1-\alpha)$, where $\hat{F}$ is the empirical CDF of $\{z_i\}$; this delivers the target FPR at any α by Glivenko–Cantelli regardless of null shape, with the Dvoretzky–Kiefer–Wolfowitz inequality giving $\sup_x|\hat{F}(x) - F(x)| = O_p(M^{-1/2})$ Monte-Carlo noise in the threshold.*

Proof. Standardise: $(z - \hat{\mu})/\hat{\sigma}$ is asymptotically $\mathcal{N}(0,1)$ as $M \to \infty$ by the standard plug-in argument (consistency of $\hat{\mu}, \hat{\sigma}$ + Slutsky), so $\Pr[z > \hat{\mu} + z_\alpha\hat{\sigma}] \to \Pr[\mathcal{N}(0,1) > z_\alpha] = \alpha$. The recipe leaves the watermarked side untouched: the watermarked z-distribution still concentrates around $\rho\sqrt{(S-1)(n-1)}$ (Theorem A.2), which on natural-language text empirically exceeds $z^\star$ by an order of magnitude (Table 8), so TPR remains at 100%. □

A.3. Universal Robustness Threshold

Theorem A.4 (Universal robustness threshold)**.** *Suppose an* oracle-blind *adversary (without access to σ_κ or the random oracle) independently replaces each token with probability δ by a fresh token whose state σ_κ is uniform on $[S]$ (so the substitutions are i.i.d. uniform in the partition; structured paraphrase / NLLB back-translation produces correlated edits and is treated empirically in Subsection 6.3). Standardise the post-attack z by the* null *standard deviation $\sqrt{p_0(1-p_0)/(n-1)}$ (the same scaling used pre-attack). Then for clockwork ChainMark with state count S and gate density ρ,*

$$\mathbb{E}[\phi \mid \text{attack}] \;=\; \tfrac{1}{S} + \rho(1-\delta)^2\tfrac{S-1}{S}. \tag{9}$$

Define the pre-attack and post-attack mean-gap signals $\Delta_{\text{pre}} = \rho(S-1)/S$ and $\Delta_{\text{post}} = \rho(1-\delta)^2(S-1)/S$. Both share the same null SD, so the standardised-margin ratio is

$$\frac{z_{\text{post}}}{z_{\text{pre}}} \;=\; \frac{\Delta_{\text{post}}}{\Delta_{\text{pre}}} \;=\; (1-\delta)^2, \tag{10}$$

independent of (S, ρ, n). Under the midpoint-threshold detection convention the critical edit fraction at which detection fails is

$$\delta^\star \;=\; 1 - 1/\sqrt{2} \;\approx\; 0.293, \tag{11}$$

again independent of (S, ρ, n).

Proof. Let $M_i \sim \text{Bernoulli}(\delta)$, i.i.d. across i, indicate that t_i has been replaced; modified tokens receive fresh uniform state by the random-oracle property. For each pair (t_i, t_{i+1}), decompose on (M_i, M_{i+1}):

- $(0,0)$: both tokens survive. The indicator equals the pre-attack distribution, with mean $1/S + \rho(S-1)/S$. Weight $(1-\delta)^2$.
- $(0,1)$: t_{i+1} fresh. By column-regularity (which is trivial for clockwork), $\mathbb{E}[X_i] = 1/S$. Weight $(1-\delta)\delta$.
- $(1,0)$: symmetric, by row-regularity. Mean $1/S$, weight $\delta(1-\delta)$.
- $(1,1)$: both fresh. Mean $1/S$, weight δ^2.

Combining,

$$\begin{aligned}\mathbb{E}[\phi \mid \text{atk}] &= (1-\delta)^2\left[\tfrac{1}{S} + \rho\tfrac{S-1}{S}\right] + (1-(1-\delta)^2)\tfrac{1}{S} \\ &= \tfrac{1}{S} + \rho(1-\delta)^2\tfrac{S-1}{S},\end{aligned}$$

proving (9).

z-ratio independence. The mean gap above the null baseline $1/S$ is $\Delta = \rho(S-1)/S$ pre-attack and $\Delta(1-\delta)^2$ post-attack. We standardise both regimes by the *null* SD $\sqrt{p_0(1-p_0)/(n-1)}$ (the conventional one-sample z-test scaling, used as a fixed denominator across attacks); this is a definition, not an assumption that the post-attack alternative variance is null-bounded. Hence $z_{\text{post}}/z_{\text{pre}} = \Delta_{\text{post}}/\Delta_{\text{pre}} = (1-\delta)^2$, independent of (S, ρ, n).

Critical fraction. At the midpoint threshold $\tau_{\text{mid}} = 1/S + (\rho/2)(S-1)/S$, detection is defeated iff the post-attack

mean drops below τ_{mid}, i.e.

$$\rho(1-\delta)^2\tfrac{S-1}{S} \;<\; \tfrac{\rho}{2}\tfrac{S-1}{S} \quad\Longleftrightarrow\quad (1-\delta)^2 \;<\; \tfrac{1}{2}.$$

Solving, $\delta^\star = 1 - 1/\sqrt{2}$. The factors $(S-1)/S$ and ρ cancel on both sides; this is the structural reason for the universal threshold. □

A.4. k-Regular Generalisation

We now lift the clockwork results to any k-regular adjacency. We recall the definition for the appendix.

Definition A.5 (k-regular topology)**.** $T \in \{0,1\}^{S\times S}$ is *k-regular* ($1 \le k \le S-1$) if every row and every column contains exactly k ones. Equivalently, T/k is doubly stochastic. The induced detection baseline is $p_0 = k/S$.

Theorem A.6 (k-regular calibration identities)**.** *Fix any k-regular adjacency T and let $p_0 = k/S$. Apply Algorithm 1 and Algorithm 2 with T in place of the clockwork matrix. Then for an i.i.d. random null sequence $\mathbf{t}^r$, a watermarked sequence $\mathbf{t}^w$ with gate density ρ, and an oracle-blind adversary with per-token replacement rate δ (as in Theorem A.4),*

$$\mathbb{E}[\phi(\mathbf{t}^r)] = p_0, \tag{12}$$
$$\mathbb{E}[\phi(\mathbf{t}^w)] = p_0 + \rho(1-p_0), \tag{13}$$
$$\mathrm{Var}[\phi(\mathbf{t}^r)] = \tfrac{p_0(1-p_0)}{n-1}, \tag{14}$$
$$z(\rho, p_0, n) = \rho\sqrt{\tfrac{(1-p_0)(n-1)}{p_0}}, \tag{15}$$
$$\mathbb{E}[\phi \mid \text{atk}] = p_0 + \rho(1-\delta)^2(1-p_0). \tag{16}$$

Proof. Equation (12) and the variance (14) are immediate from Lemma A.1.

Watermark mean. At a gated position, Algorithm 1 restricts the argmax to $\bigcup_{s'\in N^+(s)} \mathcal{V}_{s'}$, so $\sigma_\kappa(t_{i+1}) \in N^+(\sigma_\kappa(t_i))$, i.e., $X_i = 1$. At an ungated position, $X_i \sim \text{Bernoulli}(p_0)$. Averaging, $\mathbb{E}[\phi(\mathbf{t}^w)] = \rho + (1-\rho)p_0 = p_0 + \rho(1-p_0)$.

z-score. Dividing the signed mean gap $\rho(1-p_0)$ by the null SD $\sqrt{p_0(1-p_0)/(n-1)}$ gives $\rho\sqrt{(1-p_0)(n-1)/p_0}$. At $k=1$, $p_0 = 1/S$ and this reduces to $\rho\sqrt{(S-1)(n-1)}$, recovering Theorem A.2.

Post-attack mean. Repeat the four-case decomposition of Theorem A.4. The $(0,1)$ case requires column-regularity to give $\mathbb{E}[X_i \mid t_{i+1} \text{ fresh}] = p_0$ (the conditional in-degree $|N^-(s')|/S$ must be p_0 uniformly in s'); the $(1,0)$ case symmetrically uses row-regularity. Hence

$$\begin{aligned}\mathbb{E}[\phi \mid \text{atk}] &= (1-\delta)^2(p_0 + \rho(1-p_0)) + (1-(1-\delta)^2)p_0 \\ &= p_0 + \rho(1-\delta)^2(1-p_0).\end{aligned}$$

□

Corollary A.7 (Universal midpoint $\delta^\star$)**.** *Under the midpoint-threshold detection convention, every k-regular ChainMark scheme with $1 \le k < S$ (so $p_0 < 1$; the degenerate case $k = S$ has every transition valid and is trivially undetectable) has critical edit fraction*

$$\delta^\star \;=\; 1 - 1/\sqrt{2},$$

independent of the topology parameters (k, S, ρ, n).

Proof. The midpoint threshold is $\tau_{\text{mid}} = p_0 + (\rho/2)(1-p_0)$. By (16), the post-attack mean drops below τ_{mid} iff

$$\rho(1-\delta)^2(1-p_0) \;<\; \tfrac{\rho}{2}(1-p_0) \iff (1-\delta)^2 \;<\; \tfrac{1}{2},$$

which is independent of (k, S, ρ, n) once $p_0 < 1$ (i.e., $k < S$). Solving gives $\delta^\star = 1 - 1/\sqrt{2}$. □

A.5. Security Against an Oracle-Blind Adversary

Theorem A.8 (Pseudorandomness under random oracle, oracle-blind adversary)**.** *Assume the secret key κ has min-entropy at least λ and $\mathcal{H}$ is modelled as a random oracle. Consider a polynomial-time adversary $\mathcal{A}$* without *access to the random oracle,* without *access to the LM (or its per-position distribution), and* without *access to a reference text drawn from the same prompt distribution. $\mathcal{A}$ has only the watermarked output $\mathbf{t}^w$ and the public scheme parameters (S, T). Then $\mathcal{A}$ achieves:*

- *(i) key-recovery advantage at most $q \cdot 2^{-\lambda}$ for a query budget q to any auxiliary oracle that depends on κ;*
- *(ii) state-prediction success at most $1/S + \text{negl}(\lambda)$ on each fresh token $t^\star$ that has not been queried with the correct key, hence advantage $\text{negl}(\lambda)$ over the uniform $1/S$ baseline;*
- *(iii) statistical distinguishing advantage at most $\text{negl}(\lambda)$ between $\mathbf{t}^w$ and a non-watermarked LM sample $\mathbf{t}^{LM}$ from the same prompt distribution, restricted to statistics that are measurable in the partition structure σ_κ (i.e., to tests that look at state-transition patterns rather than raw token-id statistics).*

Proof. Without κ, the adversary's view of $\mathcal{H}(\kappa \,\|\, t)$ for any t is uniform on the oracle output space. Recovery probability $q \cdot 2^{-\lambda}$ follows from a standard guessing argument over a min-entropy-λ key. State prediction with advantage greater than $\text{negl}(\lambda)$ would imply distinguishing the random-oracle output from uniform, contradicting (i).

For (iii): a σ_κ-measurable statistic $\mathcal{D}(\sigma_\kappa(\mathbf{t}))$ that distinguishes $\mathbf{t}^w$ from $\mathbf{t}^{LM}$ with advantage ε must, by the random-oracle property, distinguish the watermarked state sequence (which walks the chain at gated positions) from a state sequence drawn uniformly on $[S]^n$ (the LM's σ_κ-image is uniform iid by (ii)). Such a $\mathcal{D}$ yields a state-predictor with advantage at least ε on some fresh token, contradicting (ii) by a hybrid argument; hence $\varepsilon \le \text{negl}(\lambda)$. *Token-id-level*

statistics that exploit natural-language marginals (e.g., bigram frequencies) are explicitly outside this guarantee: they are not σ_κ-measurable and trivially distinguish any LM sample from i.i.d. uniform tokens. □

A.6. Optimal LM-Aware Detector

Theorem A.9 (Entropy-weighted detector is asymptotically locally most powerful)**.** *Suppose the verifier has access to the LM at each position, and let $g_i \in \{0,1\}$ denote the gate indicator at position i. Set $\pi_i = g_i + (1-g_i)/S$, the marginal probability of $X_i = 1$ under the watermark alternative. Approximating the joint law $\{X_i\}$ by the product of its marginals (asymptotically valid under the random oracle on distinct tokens, with the product-form CLT covariance vanishing by Lemma A.1), the log-likelihood ratio under the product law is*

$$\Lambda \;=\; \sum_{i=1}^{n-1}\Big[X_i \log(\pi_i S) + (1-X_i)\log\tfrac{1-\pi_i}{1-1/S}\Big]. \quad (17)$$

Then the test $\{\Lambda > c_\alpha\}$, with c_α the size-α critical value of Λ under the null, is asymptotically locally most powerful within the product hypothesis class — i.e. it attains the Neyman–Pearson power against the product-form approximation of the joint law, with the 1-dependence of $\{X_i\}$ contributing only $o(1)$ corrections in the LAN regime.

Proof. Under the null, $X_i \sim \mathrm{Bernoulli}(1/S)$ marginally with pairwise zero covariance (Lemma A.1). Under the watermark alternative, $X_i \sim \mathrm{Bernoulli}(\pi_i)$. The sequence $\{X_i\}$ is 1-dependent: X_i and X_{i+1} share t_{i+1}, but for $|i-j| \geq 2$, $X_i \perp X_j$ under the random-oracle model on distinct tokens.

The product-form likelihood ratio is

$$\begin{aligned}\Lambda &= \sum_i \log \frac{\pi_i^{X_i}(1-\pi_i)^{1-X_i}}{(1/S)^{X_i}(1-1/S)^{1-X_i}} \\ &= \sum_i X_i \log(\pi_i S) + (1-X_i)\log\tfrac{1-\pi_i}{1-1/S},\end{aligned}$$

matching the theorem statement (with the convention $0\log 0 = 0$ for terms with $\pi_i = 1$, i.e., gated positions where $X_i = 1$ is deterministic). The Neyman–Pearson lemma applied to the product hypothesis identifies Λ as uniformly most powerful for the product law.

By the central limit theorem for 1-dependent sequences (and vanishing pairwise covariance from Lemma A.1), the product-form Λ has the same Gaussian limit as the product-form joint LLR up to $o(1)$ corrections (LAN regime, Le Cam); we therefore claim asymptotic local most-powerfulness among tests within the product hypothesis class, not strict Neyman–Pearson optimality against the true 1-dependent joint. Hence the test based on Λ achieves the Neyman–Pearson power asymptotically. □

A.7. Detector Pseudocode

We restate the model-free detection algorithm for self-contained reading; the body version is Algorithm 2.

Algorithm 3 ChainMark Watermark Detection (model-free)

Input: text x, key κ, states S, transition T, threshold τ
$\mathbf{t} \leftarrow \mathrm{Tokenize}(x)$; $c \leftarrow 0$; $n \leftarrow |\mathbf{t}|$
for $i = 1$ **to** $n-1$ **do**
  $s_i \leftarrow \mathcal{H}(\kappa \,\|\, t_i) \bmod S$; $s_{i+1} \leftarrow \mathcal{H}(\kappa \,\|\, t_{i+1}) \bmod S$
  if $T[s_i, s_{i+1}] > 0$ **then**
    $c \leftarrow c + 1$
  end if
end for
$\phi \leftarrow c/(n-1)$; $p_0 \leftarrow (\sum_{s,s'} T[s,s'])/S^2$
$z \leftarrow (\phi - p_0)/\sqrt{p_0(1-p_0)/(n-1)}$
return $(\mathbf{1}[\phi > \tau],\ \phi,\ z,\ 1-\Phi(z))$

The runtime is $O(n)$ hash evaluations and a single pass over the token stream; no LM access is required. The transition T is the same matrix used at generation time, so the random baseline p_0 is computed directly from T (clockwork: $p_0 = 1/S$; soft-cycle: $p_0 = 2/S$; general k-regular: $p_0 = k/S$).

A.8. Supplementary Results: Quality Cost and Self-Healing

These two results were stated in earlier drafts; we retain the formal statements and proofs in the appendix for completeness, since they are referenced from elsewhere in the paper.

Theorem A.10 (Quality cost identity and Jensen lower bound)**.** *At a gated position with current state s and partition mass $Z_s = p(\mathcal{V}_{(s+1) \bmod S}) > 0$, the KL between the renormalised ChainMark distribution $P_{\mathrm{ChainMark}}(t) = p(t)/Z_s \cdot \mathbf{1}[t \in \mathcal{V}_{(s+1) \bmod S}]$ and the LM distribution p is $D_{\mathrm{KL}}(P_{\mathrm{ChainMark}} \,\|\, p) = \log(1/Z_s)$. Under hash uniformity, $\mathbb{E}[Z_s] = 1/S$, and Jensen's inequality gives*

$$\mathbb{E}[D_{\mathrm{KL}}] \;\geq\; \log S, \quad (18)$$

with equality iff Z_s is constant in the key. Averaged over positions with gate rate ρ, $\mathbb{E}[D_{\mathrm{KL}}]_{\mathrm{tok}} \geq \rho \log S$.

Proof. Direct computation: $D_{\mathrm{KL}}(P_{\mathrm{ChainMark}} \,\|\, p) = \sum_{t \in \mathcal{V}_{(s+1) \bmod S}} (p(t)/Z_s)\log(1/Z_s) = \log(1/Z_s)$. Under hash uniformity, $\Pr[t \in \mathcal{V}_{(s+1) \bmod S}] = 1/S$ for every fixed t, so $\mathbb{E}[Z_s] = \sum_t p(t)/S = 1/S$. Concavity of log and Jensen give $\mathbb{E}[\log Z_s] \leq \log \mathbb{E}[Z_s] = -\log S$, hence $\mathbb{E}[\log(1/Z_s)] \geq \log S$. The token-averaged bound follows by averaging over positions with gate rate ρ. □

Theorem A.11 (Self-healing against oracle-blind adversaries)**.** *Let $\mathcal{T}_{\mathrm{q}} \subseteq \mathcal{V}$ be the set of tokens the adversary has queried with the correct key. For any token $t' \notin \mathcal{T}_{\mathrm{q}}$ that the adversary introduces, $\sigma_\kappa(t')$ is uniform on $[S]$ by the random-oracle property, so each modified position contributes an indicator distributed as* $\mathrm{Bernoulli}(1/S)$ *to the validity count, independently of strategy. Hence under an oracle-blind adversary (every introduced token unqueried), the expected mass contributed to ϕ by modified-pair positions equals*

$$\Sigma(\delta) \;=\; \tfrac{1}{S}\,\delta(2-\delta); \tag{19}$$

this is a lower bound when ranging over query-aided strategies that may bias replacement tokens toward $\sigma_\kappa(\mathcal{T}_{\mathrm{q}})$.

Proof. The fraction of token pairs (t_i, t_{i+1}) touching at least one modified token is $1-(1-\delta)^2 = \delta(2-\delta)$. Each such pair contributes an indicator with conditional expectation $1/S$ by the same column- and row-regularity argument as in Theorem A.4. Strategies in which the adversary introduces previously-queried tokens (whose state is in $\sigma_\kappa(\mathcal{T}_{\mathrm{q}})$) can craft pairs that hit valid transitions deterministically, raising ϕ above $\Sigma(\delta)$, which only helps detection. The lower bound $\Sigma(\delta) \geq \delta(2-\delta)/S$ is therefore strategy-free. □

B. Reference Tables

B.1. Closed-Form Calibration Lookup

Table 4. Calibration lookup $S^\star(n, \rho, \alpha)$ from Theorem A.2. Rows are text lengths; columns are (α, ρ) pairs. Entries are derived analytically from the closed-form detection bound and do not require empirical anchoring.

	$\alpha = 10^{-3}$		$\alpha = 10^{-6}$	
n	ρ=0.3	ρ=0.5	ρ=0.3	ρ=0.5
100	6	3	12	5
200	4	2	7	3
500	2	2	4	2
1000	2	2	3	2

C. Experimental Protocol and Reproducibility

This appendix records every setting needed to reproduce the head-to-head matched-budget study of Section 6, the per-cell tables in the main results, and every figure.

C.1. Models, Tokenizers, and Hardware

Model fleet. The headline experiments use three instruction-tuned 7–8B parameter open-weight LLMs, each loaded via the Hugging Face `transformers` library. The exact `repo_id`, revision pin, and `torch_dtype` for every model are listed in the public code release alongside its prompt template; we do not enumerate them here in order to keep the discussion model-agnostic. Each model is run on a single NVIDIA GPU (H100 or H200, depending on memory pressure); CPU fallback is supported for the detector but not for generation. Every model uses its own native tokenizer, both for generation and for detection, and self-perplexity (PPL) is computed on the same model's logits that produced the text.

Decoding. Watermarked positions use greedy argmax over the allowed-partition mask of Algorithm 1; ungated positions use temperature sampling at $T = 0.7$. The same temperature is used to draw pilot generations for gate-threshold calibration.

Gated models. A subset of the fleet is gated on the Hugging Face Hub. The released environment expects an `HF_TOKEN` in scope at runtime; the token is read by `huggingface_hub` and never logged to disk. We do not name specific gated repositories in this appendix; the public code drop records each `repo_id` alongside its license terms.

C.2. Domains and Prompts

Four domains, $N = 200$ prompts each. Each domain runs on $n = 200$ prompts drawn from a fixed pool, identical across gates, models, and attacks within a cell.

- **Code:** HumanEval Python signatures with their natural-language docstrings as the prompt; the generation is the function body.
- **Factual:** short closed-answer prompts asking for a single attested fact (capitals, dates, named entities).
- **Wiki:** open-ended descriptive prompts derived from a curated Wikipedia concept list, expanded via a fixed template.
- **Writing:** open-ended creative-writing prompts requesting a short narrative or argumentative passage.

Every prompt set, with template strings, prompt indices, and a deterministic shuffle seed, is shipped under `data/v7_min/<domain>/records.jsonl`.

C.3. ChainMark Hyperparameters

Default cell. $S = 5$ states, clockwork transition $T(s, s') = \mathbb{1}[s' \equiv s{+}1 \pmod S]$, watermark budget $\rho = 0.5$, secret key fixed across all cells, runs, and models.

State sweep. Subsection 6.5 sweeps $S \in \{2, 3, 5\}$ at fixed $\rho = 0.5$ to anchor the closed-form calibration of Theorem A.2 on data.

k-regular topology. Subsection 6.4 runs the soft-cycle ($k = 2$) topology at the default cell and compares the empirical robustness threshold against $\delta^\star$ to validate Theorem A.6.

C.4. Baseline Calibration

The two baselines, KGW (Kirchenbauer et al., 2023) and SWEET (Lee et al., 2024), are matched to ChainMark's budget at the cell level. SWEET is an entropy gate that activates the watermark only at *high*-entropy positions (the structural opposite of schemes that gate at low-entropy positions); we configure it so that the realised gate rate over the pilot pool is within ± 0.02 of $\rho = 0.5$, by quantile-matching τ_H to the $(1 - \rho)$ quantile of the per-position entropy distribution. KGW runs at the same effective budget by construction, since it gates every position. All three schemes share the same secret key and the same generation pool.

C.5. Attack Protocol

Each watermarked generation is subjected to three independent attack streams; post-attack detection statistics are reported per cell in Section 6.

- **Random substitution** ($\delta = 0.20$): $\lceil 0.20\, n \rceil$ token positions chosen uniformly at random and replaced with uniform draws from the model's tokenizer vocabulary.

- **Translation round-trip:** EN $\to$ L $\to$ EN via the NLLB-200 distilled model (NLLB Team, 2022), where $L \in \{\text{French}, \text{German}, \text{Russian}, \text{Chinese}\}$. Each language constitutes a separate cell.

We do not include grammar-preserving paraphrase attacks in this release; their evaluation is deferred to the extended version.

C.6. Randomisation

`random`, `numpy.random`, and `torch.manual_seed` are all seeded to 42 at the start of each generation run. The random-substitution attack uses an independent seed (43). Bootstrap resamples in Section 6 use a per-contrast key derived deterministically from the cell descriptor.

C.7. Released Artefacts

For each generation we record the prompt, the watermarked output, the per-position gate signal, the realised gate rate $\bar{\rho}$, PPL, the detector statistic ϕ and its z-score, and every post-attack ϕ, z, and detection flag. All tables and figures in this paper derive from these records. The code drop ships under the same repository as this paper; the per-domain JSONL records under `data/v7_min/<domain>/records.jsonl` are the source-of-truth for every empirical claim.

D. Additional Run Data and Open Extensions

D.1. Translation Per-Pivot Breakdown

Table 5. **Per-pivot ChainMark translation robustness.** TPR @ $z > 2.326$ for each NLLB-200 pivot under ChainMark $S = 5$, $\rho = 0.5$. All three models now cover four domains (factual, wiki, writing, code) at 25 prompts each ($n = 100$ per row, 300 generations per pivot column). ZH is the most aggressive pivot ($\delta_{\text{eff}} \approx 0.81$); FR, DE, RU are gentler ($\delta_{\text{eff}} \approx 0.69$).

Model	n	clean	FR	DE	RU	ZH
Llama-3.1-8B	100	100%	85%	78%	81%	75%
Mistral-7B	100	100%	86%	92%	92%	77%
Qwen-2.5-7B	100	100%	78%	75%	78%	66%

D.2. FPR Recalibration: SD Recipe and Failure Modes

Table 6. **Empirical-SD recalibration vs. failure modes, evaluated on $n = 3000$ non-watermarked samples (1000 per model).** Goal: bring empirical FPR close to the $\alpha = 1\%$ target while preserving TPR$= 100\%$. Only the empirical-SD recipe (Fix 1) brings FPR within ~ 0.2 pp of the target (1.17%); the other four sit at 2–5% FPR or collapse TPR. The iid-baseline row is the closed-form $z_\alpha = 2.326$ threshold without recalibration. The 2.07% baseline on this $n = 3000$ pooled corpus supersedes the 1.7–2.0% per-S cells in Table 3 (which use the smaller Exp. 5 sub-design with $n \approx 300$ per cell).

Method	FPR	TPR
iid baseline (no fix; $z > 2.326$)	2.07%	100.0%
Fix 1: empirical SD recal.	**1.17%**	**100.0%**
Failure modes (reported, not adopted):		
Fix 2: Newey–West HAC	3.45%	100.0%
Fix 3: k-skip ($k = 3$)	2.46%	2.2%
Fix 4: stopword filter (top-200)	5.18%	99.0%
Fix 5: HDD-lite (inv.-freq. weighted)	2.35%	100.0%

D.3. Apples-to-Apples Head-to-Head at Empirical FPR$= 1\%$

D.4. Empirical FPR$= 0\%$ Achievability

Open data and deferred extensions. The following experiments are deferred to the extended version of this paper:

- **DIPPER paraphrase attack.** Grammar-preserving paraphrase via DIPPER (Krishna et al., 2023) on at least one (model, domain) cell.

- **k-regular validation at $k \geq 3$.** Theorem A.6 predicts $p_0 = k/S$ for any k-regular topology; we validate $k \in \{1, 2\}$ only.

- **External-LM PPL or MAUVE quality metric.** Self-PPL is conservative but circular; an external judge LM or MAUVE would give a model-independent quality reading.

Table 7. **Empirically-calibrated head-to-head at FPR$= 1\%$.** Each method's threshold is the maximum across the per-model 99% z-quantiles on a non-watermarked calibration corpus (the conservative recipe; per-model nulls in ChainMark: Llama 2.87, Qwen 2.87, Mistral 2.41, so ChainMark $z^\star = 2.87$; KGW per-model: Llama 2.77, Mistral 2.34, Qwen 2.62, so KGW $z^\star = 2.77$). ChainMark retains its TPR advantage at matched empirical FPR, confirming the analytical-threshold comparison in Table 1 is not an FPR artefact. SWEET is omitted from this table: the SWEET null z-distribution was not collected on the FPR corpus and cannot be reproduced from the released artefacts; a SWEET-recalibrated row is deferred to the extended version.

Method	$\text{TPR}_{\text{clean}}$	$\text{TPR}_{\text{random}}$	TPR_{ZH}
ChainMark (ours, S=5)	**100.0%**	**99.9%**	**68.3%**
KGW $\gamma = 0.5$	66.8%	42.1%	13.0%

Table 8. **FPR$= 0\%$ by lifting the threshold to $z^\star = \max_i z_{\text{nwm},i} + 0.5$.** Computed on the held-out non-watermarked corpus ($\sim$ 1000 samples per model). Watermarked text retains TPR$= 100\%$ because clean ChainMark z at the canonical $S = 5$ ($\approx$ 25 across the three models; Table 3) is several times the maximum observed null z. KGW and SWEET cannot achieve this regime because their watermarked and non-watermarked z-distributions overlap.

Model	$\max z_{\text{nwm}}$	lifted $z^\star$	TPR @ $z^\star$
Llama-3.1-8B	7.83	8.33	100%
Qwen-2.5-7B	5.00	5.50	100%
Mistral-7B	4.82	5.32	100%

- **Base (non-instruction-tuned) and $\geq$ 10B LLMs.** Our fleet covers only instruction-tuned 7–8B models.
- **ρ-sweep on the new fleet.** The current headline runs use $\rho = 0.5$ only; the previous gate-invariance ablation was on a smaller setup.
- **Adversary with detector-oracle access.** Theorem A.8 excludes adversaries who can query the detector; characterising query complexity is open.

E. Reference Detector Implementation

The following 24-line Python listing is a self-contained, dependency-minimal reference implementation of Algorithm 2 for clockwork ChainMark. It takes a list of integer token IDs, a bytes key, a state count S, and a significance threshold α; it returns the fingerprint score ϕ, the z-score, the one-sided p-value, and a binary detection flag.

The listing is sufficient to independently verify any ChainMark-watermarked text given only the secret key, with no language-model access and no learned components. Extensions to arbitrary k-regular topologies (cf. Theorem A.6) require replacing the successor check `states[i+1] == (states[i]+1) % S` with a lookup `T[states[i]][states[i+1]] > 0` and setting `p0 = (T > 0).mean()`; no other modification is needed.

Complexity. Two SHA-256 evaluations per token ($2n{\cdot}256$ bits of hash output) and $n - 1$ integer equality checks. At $n = 100$ the entire detection pipeline runs in under 1 ms on a single CPU core; no GPU, no model weights, no tokenizer aside from what is needed to obtain `token_ids`.

Interpretation as an audit primitive. The detector's inputs are exactly what a third-party auditor could receive under an Article 50 disclosure regime: a public text, a detector program, and a (confidentially held) key. The output is a p-value with analytically known false-positive behaviour under the null, sidestepping the calibration-by-grid-search problem that afflicts logit-bias watermark families.

Key management. The scheme's security reduces to the confidentiality of `key` plus the random-oracle idealisation of SHA-256 (Theorem A.8). In practice an auditor and a deployer can share the key through any standard key-management substrate (e.g., HKDF-derived per-deployment keys, committed to an external ledger so that the commitment precedes generation). Rotating keys per deployment epoch preserves the $p_0 = k/S$ null baseline without changing any detector behaviour.

F. Prompt and Generation Examples

This appendix shows representative prompts from each of the four content domains used in Section 6, together with the exact ChainMark configuration and post-attack pipeline that produced the headline numbers in Table 1 and Figure 4. All three models receive the same prompt verbatim with no system prompt prepended; decoding parameters (T=0.7, top-p=1, n=200 tokens, deterministic prompt-order seed) are also held fixed across models and methods.

Prompt examples per domain.

- **Code** (HumanEval problem stems, 164 problems available; we use the first 100):

```
from typing import List
def has_close_elements(numbers:
List[float], threshold:  float) ->
bool:
"""Check if in given list of
numbers, are any two numbers
closer to each other than given
threshold."""
```

- **Factual** (short closed-form knowledge prompts):

```
The capital of France is
```

- **Wiki** (open-ended "Explain X..." prompts over a curated 176-entry concept list):

```
Explain Donald Trump in a
comprehensive way.
```

```python
import hashlib
from math import sqrt
from scipy.stats import norm

def sha_state(key: bytes, token_id: int, S: int) -> int:
    """SHA-256-based state assignment sigma_kappa(t) = H(kappa || t) mod S."""
    h = hashlib.sha256(key + token_id.to_bytes(8, "big")).digest()
    return int.from_bytes(h[:8], "big") % S

def detect_chainmark(token_ids, key: bytes, S: int,
                     alpha: float = 0.01) -> dict:
    """Clockwork ChainMark detector; O(n) in the number of tokens."""
    n = len(token_ids)
    if n < 2:
        return {"phi": 0.0, "z": 0.0, "p_value": 1.0,
                "is_watermarked": False}
    states = [sha_state(key, t, S) for t in token_ids]
    valid = sum(1 for i in range(n - 1)
                if states[i + 1] == (states[i] + 1) % S)
    phi = valid / (n - 1)
    p0 = 1.0 / S
    se = sqrt(p0 * (1.0 - p0) / (n - 1))
    z = (phi - p0) / se
    p_value = 1.0 - norm.cdf(z)
    return {"phi": phi, "z": z, "p_value": p_value,
            "is_watermarked": p_value < alpha}
```

Figure 5. Reference detector for clockwork ChainMark: 24 lines of dependency-minimal Python (`hashlib` + `scipy.stats`). Inputs are the tokenizer's token IDs, the secret key, and the state count; outputs are the fingerprint ϕ, z-score, one-sided p-value, and a binary detection flag.

- **Writing** (creative-completion prompts):

```
Write a short story that begins:
The colony ship arrived three
centuries late, and someone was
already waiting.
```

ChainMark generation configuration (used in Table 1). At each generation step on the prompt above, the model produces a distribution over its native vocabulary; ChainMark then masks logits at gated positions (gate density ρ=0.5, high-entropy gate $G_{H_{\text{high}}}$) to keep only token IDs whose state $\sigma_\kappa(t) = \text{SHA-256}(\kappa \,\|\, t) \bmod S$ equals $(s+1) \bmod S$, where s is the previous token's state. At ungated positions the model samples at T=0.7 from the original distribution. The secret key κ is fixed for the full campaign; in deployment a regulator-side audit re-derives the same state sequence from the same key with the model's tokenizer.

Attack-pipeline example. Given the watermarked text x (say, 200 tokens of a wiki-domain answer), the three attack conditions in Table 1 apply respectively:

- **Clean.** Detect on x as-is.
- **Random substitution at δ_{eff}=0.20.** Replace a uniformly random 20% of tokens with a uniformly sampled vocabulary token, then detect on the corrupted text.
- **ZH back-translation.** Pass x through NLLB-200 (NLLB Team, 2022) English → Chinese → English; record the empirical token-edit-distance δ_{eff} (≈ 0.81 for ZH on average, Table 5); then detect on the round-tripped text.

The detector runs the same σ_κ and counts valid transitions; reports ϕ, the standardised z, and a binary flag at $z > z_\alpha$ (Algorithm 3, Figure 5).

Token-level walkthrough. For the wiki prompt `Explain Donald Trump in a comprehensive way.` on Llama-3.1-8B-Instruct at S=5, ρ=0.5, an illustrative excerpt of the first 20 generated tokens looks roughly as below. We show the (token, state) pair at each position and mark gated steps with • (the gate raised the mask; ChainMark forced $s_{i+1} = (s_i{+}1) \bmod 5$) versus ungated steps with ∘ (no mask; the model sampled freely):

Donald$_2$ ∘ **Trump**$_3$ ∘ (**born**$_4$ • **June**$_0$ • **14**$_1$ •, **1946**$_2$ •) **is**$_3$ • **a**$_4$ ∘ **former**$_4$ ∘ **U.S.**$_0$ • **president**$_1$ • **and**$_2$ •

`businessman`$_3$ • ...

At gated positions the next token's state is forced to be $(\text{prev} + 1) \bmod 5$ (here $S{=}5$, so the cycle is $0{\to}1{\to}2{\to}3{\to}4{\to}0$); the model picks the highest-prob *token* whose state matches that requirement. At ungated positions any token may follow. The gate threshold τ is calibrated on a pilot so the realised gate rate tracks $\rho{=}0.5$.

Detection on this excerpt. A regulator running the detector (Figure 5) on the same string with the same key κ re-derives the state sequence $2, 3, 4, 0, 1, 2, 3, 4, 4, 4, 0, 1, 2, 3, \ldots$ and observes $11/14$ valid transitions, i.e. $\phi = 0.79$. With $p_0{=}0.20$ and $n{=}14$, $z = (0.79 - 0.20)/\sqrt{0.16/13} \approx 5.3$, well above $z_{0.01} = 2.326$, so the detector flags "watermarked". A non-watermarked sample of the same length under the null distribution averages $\phi \approx 0.20$ and $z \approx 0$.