

The Provenance Horizon: Short Regulated Text Resists Reliable Watermarking Under Editing

Yanmin Zhao¹ Jikun Wu² Tengjiao Zheng³ Yang Liu³
Dongxin Guo² Siu Ming Yiu^{1*}

¹The University of Hong Kong ²Brain Investing Limited ³Stellaris AI Limited
{ymzhao, smyiu}@cs.hku.hk
{jacobwu, bettyguo}@brainvestai.com
{tomzheng, yangliu}@stellaris-ai.com

Abstract

The EU Artificial Intelligence Act now requires providers to mark generated text so its origin can be checked. In regulated industries the duty is sharper: a marker must never flag third-party writing, survive routine review edits, and work on outputs often a sentence long. We call these the regulated-provenance triad and prove that no bounded-distortion watermark perturbing a model’s token distribution can satisfy all three below a critical length, the provenance horizon. The proof combines the Le Cam two-point method with a strong data-processing inequality for the edit channel; a matching achievability bound for the green-list family under a memoryless resampling channel makes the scaling order-tight. The barrier rises at negligible-distortion budgets, not at deployed ones. The contraction factor entering the bound is a proven upper bound for the memoryless resampling channel and a calibrated estimate for realistic paraphrase, under which we estimate most sentence-length regulated outputs fall below the horizon. Relocating provenance to a tamper-evident log read through a conformally calibrated retrieval test sidesteps the barrier: under exchangeable calibration its false-positive rate is at or below the target at every length. Coverage carries no guarantee and degrades under an adaptive editor.

1 Introduction

Large language models now generate fluent text at scale (Brown et al., 2020), and provenance for it has moved from a research curiosity to a legal requirement. Article 50 of the EU Artificial Intelligence Act, whose transparency obligations took effect August 2026 (with a machine-readable marking transition to December 2026 under the Digital Omnibus amendment (European Parliament and Council of the European Union, 2026)), obliges providers to mark text outputs in a machine-readable, detectable

way (European Parliament and Council of the European Union, 2024; European Commission, 2026), and the accompanying Code of Practice concludes that no single marking technique is sufficient in all cases to meet the Article 50(2) standard (European Commission, AI Office, 2026). In a regulated industry the obligation is sharper: a broker-dealer drafting client communications, disclosures, and filings with a model must preserve and supervise them under SEC Rule 17a-4 and FINRA Rule 3110 (U.S. Securities and Exchange Commission, 2023; Financial Industry Regulatory Authority, 2014). A provenance mechanism here must satisfy three properties at once: not flag a client’s or counterparty’s own writing as model output, so its false-positive rate sits below a regulator-grade bound α ; still fire after the redaction, disclaimer insertion, and light paraphrase that supervisory review performs; and cover the output even when it is a single recommendation or risk sentence. We name this joint requirement the *regulated-provenance triad*.

Token-level watermarking is the default answer: green-list biasing (Kirchenbauer et al., 2023), Gumbel sampling (Aaronson, 2023), distortion-free schemes (Kuditipudi et al., 2024), and production systems such as SynthID-Text (Dathathri et al., 2024) embed a statistical signal during generation, evaluated against quality and generic robustness (Piet et al., 2025) but never against the conjunction regulation imposes: a hard false-positive bound, a bounded-edit class, and a coverage floor on short outputs. Whether a watermark can meet that conjunction is the question a compliance officer, and soon an auditor, will ask.

The difficulty is information-theoretic. A watermark is detectable only in proportion to the statistical distance it injects, which grows with output length times the per-token distortion it may add; quality caps that distortion at a budget γ , and permitted edits contract whatever survives by a factor η . A long document accumulates enough

* Corresponding author.

evidence; a one-sentence disclosure does not. The three requirements pull against one another precisely where the mandate bites: short regulated text under editing.

Existing negative results do not reach this regime: Zhang et al. (2024) assume a quality-preserving rewriter with no edit-distance bound, and Sadasivan et al. (2025) bound detection by human-machine TV distance; neither addresses a *bounded* edit budget with *nonzero* distortion and a coverage floor. The statistical detection theory (Huang et al., 2023; Li et al., 2025; He et al., 2026) establishes clean-channel token-budget scaling but does not couple it to a regulatory FPR bound, a compliance edit channel, and a short-output coverage floor. As of August 2026, we are not aware of the regulated triad being posed as a single joint constraint.

We formalize the triad as a binary hypothesis test between edited third-party and edited watermarked text (Section 3) and prove a length-threshold impossibility (Section 4): no token-distribution-perturbing watermark with budget γ , through an edit channel with KL contraction η , can satisfy it below the *provenance horizon* $L^*(\alpha, \beta, \gamma, \eta) = \frac{1}{\eta\gamma} \ln \frac{1}{1-(1-\alpha-\beta)^2}$. Figure 1 contrasts the two threat models the paper keeps separate: a stateless, third-party-verifiable watermark (T1) and a provider-side, log-holding test (T2). The impossibility is a general statement about short text under bounded edits; finance supplies the regulatory bounds and length distribution that make it concrete, not a hypothesis of the theorem.

Contributions.

- We introduce the **regulated-provenance triad**: provenance coupling a regulator-grade false-positive bound, robustness to a bounded permitted-edit class, and a coverage floor at length L , grounded in the EU AI Act, SEC Rule 17a-4, and FINRA Rule 3110 (Section 3). The triad addresses three of the four Article 50(2) requirements (reliability, robustness, effectiveness); interoperability is orthogonal. The formalization is domain-general; finance is its first worked instance.
- We prove a **closed-form provenance horizon** (Theorem 1): no γ -bounded watermark satisfies the triad for $L < L^*$. Beyond the known clean-channel token-budget scaling we add the edit-channel contraction η via a strong data-processing inequality, a proven upper bound

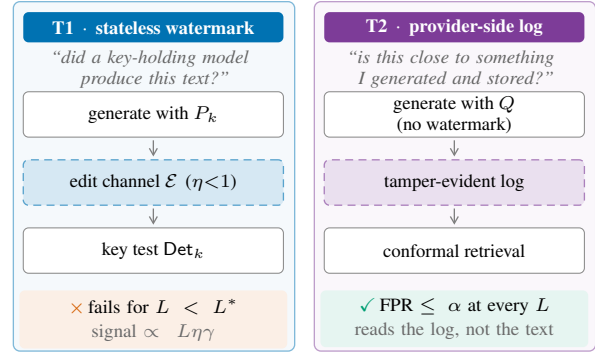


Figure 1: Two provenance threat models, answering different questions (italics) and not forming a single ranking. **T1** (stateless, third-party-verifiable): the signal lives inside the token distribution and scales as length times per-token distortion times edit survival ($L\eta\gamma$), so it vanishes below the provenance horizon L^* . **T2** (provider-side, log-holding): the signal is the regulated firm’s tamper-evident log, required under SEC Rule 17a-4, read through a conformal retrieval test; its detector reads the stored log rather than the text alone, sidestepping Theorem 1, and under exchangeable calibration its FPR is at or below the target at every length. Shaded boxes mark the contested step in each pipeline (the editing channel for T1, the stored log for T2).

when η is the green-list per-pair contraction under resampling, and an empirical surrogate under paraphrase (Section 4).

- We **prove a matching achievability bound** (Theorem 2) and **sharpen its constant** (Theorem 3): balancing the detector threshold to the two error budgets yields a computable operative length $L_{\text{op}} = \Theta(L^*)$ a small constant above L^* (Section 5).
- We **estimate η and γ** for four deployed watermarks and four edits; at the aspirational budget most sentence-level SEC EDGAR outputs fall below the horizon under illustrative light and moderate edits (95 to 98%, a calibrated estimate under surrogate η values and a lower bound since $L_{\text{feasible}} \geq L^*$), unit-dependent and largely vanishing at the filing level (Section 5).
- We **relocate provenance to the mandated log**: split-conformal calibration gives a length-, entropy-, and edit-independent false-positive guarantee (Proposition 1), which a committed log and an independent calibration set extend to an untrusting auditor (Proposition 2); an observation that sidesteps the impossibility, not a third theorem (Section 6).

2 Related Work

Token-distribution watermarks. Most deployed and proposed schemes embed a key-dependent signal by perturbing the generative distribution: green-list logit biasing (Kirchenbauer et al., 2023, 2024), Gumbel and inverse-transform sampling (Aaronson, 2023), distortion-free constructions (Kuditipudi et al., 2024; Christ et al., 2024; Hu et al., 2024), the provably robust Unigram scheme (Zhao et al., 2024), consolidations of detection, evaluation, and multi-bit watermarking (Fernandez et al., 2023), publicly verifiable variants (Fairoze et al., 2025), and the production SynthID-Text system (Dathathri et al., 2024). Logit-shift schemes share a small per-token distortion budget γ for quality; key-deterministic schemes (Gumbel, ITS) have zero *marginal* distortion but large *per-step conditional* γ , placing them outside Theorem 1’s bite (the horizon collapses at large γ). All are benchmarked on detectability and generic robustness (Piet et al., 2025), none against a regulator-grade false-positive bound jointly with a bounded-edit class and a coverage floor. Notably, Kuditipudi et al. (2024) report detection from 35 tokens under 40–50% random edits via a non-asymptotic permutation p -value, but their own low-entropy evaluation shows only $\sim 25\%$ of responses (median length ~ 100 tokens) detectable at $p \leq 0.01$, consistent with the short-output, low-entropy difficulty the triad formalizes.

Robustness limits and impossibility. Two negative results are closest. Zhang et al. (2024) prove *strong* watermarking (robustness to a quality-preserving rewriter) impossible via an attacker with black-box access to quality and perturbation oracles, and Sadasivan et al. (2025) bound any detector’s AUROC by the human–machine total-variation distance; empirical attacks corroborate both, from paraphrase evasion (Krishna et al., 2023) to watermark stealing (Jovanović et al., 2024) and zero-shot detection (Mitchell et al., 2023) defeated by the same paraphrase attacks. Our result is mechanically different: it holds for a *bounded* edit budget and *nonzero* distortion and yields a closed-form length threshold, via a data-processing contraction on the edit channel. Closest on the edit channel, Li et al. (2026) cast detection under edits as a hypothesis test and Golowich and Moitra (2024) construct watermarks robust to a constant fraction of adversarial edits in edit distance, with alphabet size and

text length polynomial in the security parameter; the result is complementary to our short-output, low-distortion regime. Chakraborty et al. (2024) note detection is possible with enough samples, consistent with our threshold being a statement about *short* outputs.

Statistical theory of watermark detection. A recent line casts detection as hypothesis testing and characterizes optimal rules and token budgets: uniformly most powerful watermarks and nearly matching bounds (Huang et al., 2023), closed-form false-negative rates and pivotal statistics (Li et al., 2025), tight fidelity-loss bounds (He et al., 2026) and detection-distortion tradeoffs (Cai et al., 2026), and debiasing via coupling (Xie et al., 2025); we build on the Neyman–Pearson machinery (Neyman and Pearson, 1933). The clean-channel token-budget scaling is established in that line (stated as a rate in entropy per token by Huang et al. 2023, who already treat a perturbation class). Our contribution over this line is the edit-channel contraction factor η that turns a clean-channel budget into an edit-robust one and couples it with all three regulatory ingredients in a single closed-form threshold; none of the prior works targets a regulatory false-positive bound under a bounded edit channel. Closest, He et al. (2026) give a tight clean-channel fidelity-loss bound under Type-I/II constraints via an f -divergence argument.

Provenance by logging, retrieval, and standards. An alternative to embedding a signal is to record outputs and check membership later. Krishna et al. (2023) propose a retrieval defense that survives paraphrase, and exact-match indices (Ferragina and Manzini, 2000), now scaled to trillion-token corpora (Liu et al., 2024; Xu et al., 2025), make corpus-scale retrieval and data auditing practical (Liu et al., 2025; Elazar et al., 2024). Provenance-by-metadata standards such as C2PA Content Credentials (Coalition for Content Provenance and Authenticity, 2024) certify history rather than detect copies. Our construction invents neither retrieval nor conformal prediction; it observes that a split-conformal layer (Vovk et al., 2005; Shafer and Vovk, 2008; Angelopoulos and Bates, 2023; Lei et al., 2018; Angelopoulos et al., 2024) on the already-mandated log turns a heuristic similarity score into a distribution-free false-positive guarantee that escapes Theorem 1. Table 5 (Appendix C) positions the work against prior entries.

3 Problem Formulation

Let Q be the joint law on $\mathcal{X}^L$ of legitimate (third-party or human) text over alphabet $\mathcal{X}$, written autoregressively as $Q(x) = \prod_t Q(x_t \mid x_{<t})$. A *watermarking scheme* $\mathcal{W} = (\text{Gen}_k, \text{Det}_k)$ with secret key k replaces generation by a perturbed law P_k , also autoregressive, and equips a detector $\text{Det}_k : \mathcal{X}^* \rightarrow \{0, 1\}$, where 1 means provenance-positive. Outputs have length L . A *permitted edit channel* $\mathcal{E}$, a Markov kernel on $\mathcal{X}^*$, models the transformations that compliance and redaction workflows, and benign third parties, may apply, and the detector adjudicates documents *after* the workflow, so it sees $\mathcal{E}$ -processed text in both branches.

Definition 1 (Calibration on edited text). $\mathcal{W}$ is α -calibrated if for third-party text $Y \sim Q$ passed through the workflow, $\Pr_{Y, \mathcal{E}}[\text{Det}_k(\mathcal{E}(Y)) = 1] \leq \alpha$.

Definition 2 (Coverage under editing). $\mathcal{W}$ is $(1 - \beta)$ -coverage-robust to $\mathcal{E}$ if for an output $X \sim P_k$, $\Pr_{X, \mathcal{E}}[\text{Det}_k(\mathcal{E}(X)) = 1] \geq 1 - \beta$.

Definition 3 (Distortion budget). $\mathcal{W}$ is γ -bounded if for every output length L the average per-step conditional divergence under the watermarked path measure P_k obeys $\frac{1}{L} \sum_{t=1}^L \mathbb{E}_{x_{<t} \sim P_k} \text{KL}(P_k(\cdot \mid x_{<t}) \parallel Q(\cdot \mid x_{<t})) \leq \gamma$.

Definition 4 (Edit contraction). $\mathcal{E}$ has KL strong-data-processing (contraction) coefficient $\eta := \eta_{\text{KL}}(\mathcal{E}) = \sup_{\mu \neq \nu} \text{KL}(\mathcal{E}\mu \parallel \mathcal{E}\nu) / \text{KL}(\mu \parallel \nu) \in (0, 1)$.

We take Q to be both the legitimate-text law and the unwatermarked generative law, the regime where the watermark carries the only usable signal; where the two differ materially, zero-shot detection is available and the triad is not the binding constraint. Appendix F estimates γ against the base model under this assumption. Calibration encodes the ban on misattributing third-party text, coverage the marking mandate, the distortion budget the quality requirement, and the contraction coefficient how much distinguishing information the edits destroy. We do *not* assume the channel fixes the legitimate-text law (disclaimer insertion changes it); the proof applies the contraction directly to the two edited laws. Definition 3 averages the per-step divergence under the watermarked path measure P_k , matching the KL chain rule $\text{KL}(P_k \parallel Q) = \sum_t \mathbb{E}_{P_k}[\cdot] \leq L\gamma$ directly. Definition 4 states the worst-case coefficient; Theorem 1 uses only the per-pair contraction

$\text{KL}(\mathcal{E} \circ P_k \parallel \mathcal{E} \circ Q) / \text{KL}(P_k \parallel Q)$ at the single pair (P_k, Q) , which is at most η and may be smaller: a proven upper bound for the resampling channel of Section 5 (Appendix E), and a plug-in estimate (Appendix F) for paraphrase. The *regulated-provenance triad* is the simultaneous satisfaction of Definitions 1–3 against a channel with per-pair contraction $\eta < 1$.

In the broker-dealer setting Q is the legitimate text the firm receives and produces, P_k its watermarked drafting assistant, $\mathcal{E}$ the supervisory-review pipeline (redaction, disclaimer insertion, light paraphrase), α the tolerable rate of flagging a client’s words as machine output, and $1 - \beta$ the fraction of machine-drafted communications it must identify.

The regulated unit. The triad is a per-output statement, so the length L that governs it is the length of the unit a regulator supervises. We take the sentence as the analysis unit because it is the smallest span that carries an independently reviewable recommendation or risk statement under FINRA Rule 3110’s supervisory review; the rule does not specify a unit or mandate review of every communication, and Table 2 reports the full sensitivity to this choice. The choice matters: the barrier is the common case at the sentence level and recedes as the unit grows to a filing (Table 2), so we report the EDGAR fraction as a band over units.

4 The Impossibility Theorem

Theorem 1 (Provenance horizon). *Let $\mathcal{W}$ be any γ -bounded token-distribution-perturbing watermark and let $\eta := \text{KL}(\mathcal{E} \circ P_k \parallel \mathcal{E} \circ Q) / \text{KL}(P_k \parallel Q)$ be the per-pair KL contraction of $\mathcal{E}$ at (P_k, Q) . If $\mathcal{W}$ is simultaneously α -calibrated and $(1 - \beta)$ -coverage-robust to $\mathcal{E}$ on outputs of length L , with $\alpha + \beta < 1$, then*

$$L \geq L^*(\alpha, \beta, \gamma, \eta) = \frac{1}{\eta\gamma} \ln \frac{1}{1 - (1 - \alpha - \beta)^2}. \quad (1)$$

Equivalently, for every $L < L^*$ no γ -bounded scheme satisfies the triad.

The per-pair ratio η is at most the worst-case coefficient $\eta_{\text{KL}}(\mathcal{E})$ of Definition 4, so the theorem is conservative relative to worst-case bounds, and key-averaging only shrinks the relevant TV distance. For the memoryless resampling channel of Section 5, η is at most $(1 - \rho)^2$ (Appendix E), a proven upper bound that makes L^* conservative

there. For paraphrase we substitute the estimated contraction of Appendix F, a plug-in surrogate for the per-pair value rather than a proven bound, so the horizon there and every output fraction derived from it are calibrated estimates, not consequences of this theorem.

Proof sketch (full proof in Appendix D). After the workflow, the detector is a test of $H_0: Z \sim \mathcal{E} \circ Q$ against $H_1: Z \sim \mathcal{E} \circ P_k$ with Type-I error $\leq \alpha$ and Type-II error $\leq \beta$. The Le Cam two-point bound gives $\alpha + \beta \geq 1 - \text{TV}(\mathcal{E} \circ Q, \mathcal{E} \circ P_k)$, so the triad forces

$$\text{TV}(\mathcal{E} \circ Q, \mathcal{E} \circ P_k) \geq 1 - \alpha - \beta. \quad (2)$$

Applying the Bretagnolle–Huber inequality (Bretagnolle and Huber, 1978; Tsybakov, 2009) and then the strong data-processing inequality (Polyanskiy and Wu, 2025) directly to the two edited laws,

$$\text{TV}(\mathcal{E} \circ Q, \mathcal{E} \circ P_k) \leq \sqrt{1 - e^{-\eta \text{KL}(P_k \| Q)}}. \quad (3)$$

The KL chain rule, with the expectation under P_k as in Definition 3, gives $\text{KL}(P_k \| Q) = \sum_{t=1}^L \mathbb{E}_{x_{<t} \sim P_k} \text{KL}(P_k(\cdot | x_{<t}) \| Q(\cdot | x_{<t})) \leq L\gamma$. Substituting into (3) and combining with (2) yields $1 - \alpha - \beta \leq \sqrt{1 - e^{-\eta L\gamma}}$, i.e. $\eta L\gamma \geq \ln \frac{1}{1 - (1 - \alpha - \beta)^2}$, which rearranges to (1). No invariance of Q under $\mathcal{E}$ is used. $\square$

Corollary 1 (Finance instantiation). *Instantiate $\mathcal{E}$ with the operations permitted under FINRA Rule 3110 supervisory review and SEC Rule 17a-4 record handling (redaction, disclaimer insertion, light paraphrase), set α to a regulator false-positive bound and $1 - \beta$ to a coverage mandate. Then any γ -bounded watermark fails the triad on regulated outputs shorter than L^* . Section 5 estimates the fraction of SEC EDGAR sentence-level outputs shorter than L^* under a surrogate η .*

Corollary 1 substitutes regulatory constants into Theorem 1; it is a worked instantiation rather than an independent result, and the impossibility itself is domain-agnostic. The numerical operating points ($\alpha=10^{-2}$, $1-\beta=0.95$, the sentence as the regulated unit) are illustrative design choices motivated by FINRA Rule 3110 (supervisory review) and SEC Rule 17a-4 (retention), not values mandated by those rules; Table 1 ranges over a broad grid to show sensitivity.

The bound is driven by the product $\eta\gamma$ through the KL chain rule: the impossibility is specifically about token-distribution-perturbing watermarks with $\gamma > 0$. The construction of Section 6

escapes the theorem not by setting $\gamma = 0$ within the same hypothesis pair, but by changing the detection mechanism entirely: its detector reads a stored log rather than distinguishing $\mathcal{E} \circ P_k$ from $\mathcal{E} \circ Q$ on the text alone, so the two-point reduction in Step 1 of the proof does not apply.

Corollary 2 (Quality is the enemy of provenance). *Because $L^* \propto 1/\gamma$, the very property that makes a watermark deployable, a small quality footprint (small γ), strictly raises the horizon. Quality and short-output coverage cannot co-hold below L^* : reducing γ pushes L^* up and shrinks the set of coverable lengths. The contraction η is fixed by the permitted edits, not by the watermark, and at fixed β , tightening α has a comparatively weak effect (Figure 2).*

Remark 1 (A worked example). With $\alpha = 10^{-2}$, $1 - \beta = 0.95$, $\gamma = 0.02$, and moderate paraphrase $\eta = 0.5$, (1) gives $L^* \approx 216$ tokens, far above a sentence-length disclosure; at the same β , $\alpha = 10^{-3}$ gives ≈ 231 (Figure 2), and aggressive $\eta = 0.2$ exceeds 500 (Table 1).

5 Matching Bound and Computed Frontier

A matching achievability bound. Theorem 1 is a lower bound: below L^* nothing works. To show the horizon is order-tight in (η, γ) and not merely conservative, we prove a matching upper bound for a concrete, analytically clean instance of Definitions 1–4; the operative length is then a scheme-specific constant times L^* , as the simulation below quantifies. Define the *token-resampling channel* R_ρ : each token is, independently, replaced by a fresh draw from Q with probability ρ and kept with probability $1 - \rho$. For the green-list family with green fraction $g = \frac{1}{2}$, the post-edit green rate is $g + (1 - \rho)\delta$ where δ is the per-token bias, so the distinguishing signal attenuates by $1 - \rho$ and the per-pair KL contraction ratio at (P_k, Q) is at most $(1 - \rho)^2$ (Appendix E; the worst-case $\eta_{\text{KL}}(R_\rho)$ over all pairs is at most $1 - \rho$ by joint convexity, but Theorem 1’s proof uses only the per-pair value, so the tighter $(1 - \rho)^2$ applies).

Theorem 2 (Achievability). *For the channel R_ρ and a green-list scheme with $g = \frac{1}{2}$ and per-token budget γ , let $\delta^*(\gamma)$ be the unique positive solution of $\text{KL}(\text{Bern}(g + \delta^*) \| \text{Bern}(g)) = \gamma$. The green-count test with a midpoint threshold (set using ρ) achieves α -calibration and $(1 - \beta)$ -coverage for*

every $L \geq L_{\text{ach}}$, where

$$L_{\text{ach}} = \frac{2}{\eta \delta^{*2}(\gamma)} \ln \frac{1}{\min(\alpha, \beta)}. \quad (4)$$

Since $\delta^{*2} = (\gamma/2)(1 - \gamma/3 + O(\gamma^2))$ at $g=\frac{1}{2}$, this equals $\frac{4}{\eta\gamma} \ln \frac{1}{\min(\alpha, \beta)} \cdot (1 + O(\gamma))$, a correction of 0.67% at $\gamma=0.02$ and under 10% at $\gamma=0.24$. Together with Theorem 1, $L^* \leq L_{\text{feasible}} \leq L_{\text{ach}}$, and both are $\Theta(1/(\eta\gamma))$, so $L_{\text{feasible}} = \Theta(L^*)$ for this channel and scheme family. The result is exact for the context-independent (Unigram) green partition; for hash-based schemes with context width ≥ 1 , the tensorization and green-rate homogeneity are standard idealizations (Appendix E).

The proof (Appendix E) applies a non-asymptotic Hoeffding bound on the green-count statistic using the exact bias δ^* . The constant is loose ($L_{\text{ach}}/L^* \approx 8.6$ at $(\alpha, \beta, \gamma) = (10^{-2}, 0.05, 0.02)$), while the optimal Neyman–Pearson test does much better (simulation below). The $\Theta(1/(\eta\gamma))$ scaling is exact regardless of the correction. The matching result is limited to the green-list family under the memoryless resampling channel.

A sharper operative length. The midpoint bound (4) charges both errors the worse budget $\min(\alpha, \beta)$. Balancing the threshold to the two budgets sharpens it to a computable operative length $L_{\text{op}} = (\sqrt{\ln(1/\alpha)} + \sqrt{\ln(1/\beta)})^2 / (2\eta \delta^{*2}(\gamma))$ (Theorem 3, Appendix E), a strict improvement whenever $\alpha \neq \beta$; with Theorem 1, $L^* \leq L_{\text{feasible}} \leq L_{\text{op}}$, and the optimal Neyman–Pearson test attains the still-smaller Gaussian threshold L_{CLT} . The operative feasible length is thus a small constant times L^* , about 7 for L_{op} and 3.7 for the single-variance $\gamma \rightarrow 0$ Gaussian threshold at $(\alpha, \beta, \gamma) = (10^{-2}, 0.05, 0.02)$ against 8.6 for the midpoint; the split-threshold bound is proved by Hoeffding rather than a Gaussian approximation, so L_{op} is a genuine guarantee (Figure 3, Appendix G).

The frontier. Table 1 ranges L^* from 81 to 956 tokens over the (α, β) grid at $\gamma = 0.02$; edit strength dominates ($L^* \propto 1/\eta$), and halving γ doubles it (Corollary 2).

Estimating η and γ . The horizon bites only at the (η, γ) real schemes and edits realize, so we estimate both (Table 6; Appendix F): γ as the mean per-token KL between watermarked and base logits, η as the post-edit/pre-edit KL ratio of the detection statistic. The logit-shift schemes run at

regime (γ, η)	$\alpha=10^{-3}$ $\beta=.01$	$\alpha=10^{-2}$ $\beta=.05$	$\alpha=.05$ $\beta=.10$
light (.02, .8)	239	135	81
moderate (.02, .5)	383	216	129
aggressive (.02, .2)	956	538	321
high-quality (.01, .5)	765	431	257

Table 1: L^* (tokens, rounded up) from (1) over a regulator-relevant grid. The η values label the illustrative light/moderate/aggressive frontier; estimated η for real edits is in Table 6.

$\gamma \approx 0.19$ to 0.24 nats, roughly an order of magnitude above the aspirational 0.02 budget (chosen as an order of magnitude below deployed practice to represent the negligible-distortion regime quality-sensitive deployments target), buying detectability with quality (Corollary 2); paraphrase has η well below one (0.47 DIPPER, 0.34 GPT-4o) while light edits leave $\eta \approx 0.83$, which is why paraphrase erases detection and light edits do not. At $\gamma = 0.02$ and $\eta = 0.47$, $L^* \approx 229$. Neither estimate is a bound on the theorem’s quantity. $\hat{\eta}$ is a detection-statistic ratio whose relation to the sequence-level per-pair contraction is not established in either direction, though Table 8 indicates $\hat{\eta}$ overstates the operative contraction on every channel tested (Section 7).

Scope distinction. L^* and L_{ach} are theorem-backed for the memoryless resampling channel, where $\eta = (1 - \rho)^2$ is a proven upper bound on the per-pair contraction for the green-list family (Appendix E). All paraphrase-horizon values and EDGAR fractions are calibrated estimates based on a surrogate $\hat{\eta}$ (Appendix F), not consequences of Theorem 1.

Scaling agreement. For the resampling-channel probe (a strongly detectable green-list at $\gamma \approx 0.32$, distinct from the 0.02 frontier), the optimal-test Gaussian threshold L_{CLT} matches L_{exact} (the smallest length at which a deterministic threshold test meets both error budgets, found by exact binomial enumeration) to within 9%, while L^* sits 3.8 to 4.1 times below L_{exact} : the bound captures the $1/(\eta\gamma)$ scaling, not the constant (Appendix G, Table 7).

The barrier on EDGAR. The theorem bites at the aspirational $\gamma=0.02$, not at deployed budgets $\gamma \approx 0.19$ to 0.24 (Section 7). On the EDGAR length distribution (Appendix K), at $\gamma=0.02$ the fraction below L^* depends strongly on the regulated unit (Table 2): 95% of sentences fall below L^* under light edits and 98% under moderate edits, while at the filing level the barrier largely

Regulated unit	light ($L^*=135$)	moderate (216)	aggressive (538)
sentence (26)	95	98	>99
filing excerpt (1400)	3	6	22

Table 2: Percent of SEC EDGAR regulated outputs shorter than L^* (at $\alpha=10^{-2}$, $\gamma=0.02$), by the unit taken to be the regulated output and by edit strength. Median token counts in parentheses. The headline 98% is the sentence-level, moderate-edit cell; with the filing as the unit the barrier largely evaporates. These are calibrated estimates using illustrative η values 0.8/0.5/0.2 (Table 1) and ceiling-rounded L^* values; estimated η for real edits is in Table 6.

evaporates. The paragraph unit is omitted because paragraph-level fractions were inconsistent with the paragraph length distribution on recomputation (Appendix K); the sentence and filing rows were recomputed and are consistent. The EDGAR fraction is a band over units, the basis for the length-aware policy of Section 8.

6 Sidestepping the Horizon: A Conformal Log Test

Theorem 1 closes the door on bounded-distortion watermarks but leaves it open for mechanisms outside the token-perturbing class. We make a simple observation rather than prove a third theorem: a mechanism that already exists in the regulated setting walks through that door. In the broker-dealer setting the regulated firm already maintains a tamper-evident log of model-drafted outputs under SEC 17a-4 (U.S. Securities and Exchange Commission, 2023). Let $\mathcal{L}$ be that log and let $s(z) = \max_{r \in \mathcal{L}} \text{sim}(z, r)$ be a retrieval score, computed with an exact-match or embedding index (Krishna et al., 2023; Liu et al., 2024). Calibrate on n held-out third-party texts $\{y_i\}$ exchangeable with deployment third-party text: set $\hat{\tau}$ to the $\lceil (1 - \alpha)(n + 1) \rceil$ -th smallest of $\{s(y_i)\}$ and flag z iff $s(z) > \hat{\tau}$ (Algorithm 1, Appendix A).

Proposition 1 (Distribution-free FPR). *For any fresh third-party Y exchangeable with the calibration set, provided $n \geq \lceil 1/\alpha \rceil - 1$ and assuming continuous scores or randomized tie-breaking, $\Pr[s(Y) > \hat{\tau}] \leq \alpha$, for every output length L , every output entropy, and every edit applied to the provider’s own outputs.*

Proof. $s(Y)$ and $\{s(y_i)\}$ are exchangeable, so the rank of $s(Y)$ is uniform on $\{1, \dots, n + 1\}$ and the split-conformal threshold gives marginal coverage

$\Pr[s(Y) > \hat{\tau}] \leq \alpha$ (Vovk et al., 2005; Lei et al., 2018). The score and threshold never reference the generative law, so the guarantee does not depend on γ , L , or $\mathcal{E}$. $\square$

The guarantee above assumes the verifier trusts the provider’s log. The contested case is an auditor who does not; a binding commitment to the log (e.g. a signed Merkle root published before any query), together with an independently held and exchangeable calibration set, closes it, restoring the $\leq \alpha$ false-positive guarantee for that auditor (Proposition 2, Appendix B).

Threat model and its limits. The guarantee evades Theorem 1 because the retrieval detector uses side information (the stored log) rather than distinguishing $\mathcal{E} \circ P_k$ from $\mathcal{E} \circ Q$ on the text alone; the two-point reduction underlying the theorem does not apply to a detector with access to the generation record. We state the trade plainly. The watermark of Section 3 is stateless and third-party-verifiable; the construction is provider-side and log-holding, answering “is this close to something I generated and stored,” not “did a key-holding model produce this.” The guarantee is one-sided ($\leq \alpha$ FPR, no coverage guarantee) and marginal: an adversary crafting text near a logged output, or the boilerplate subpopulation of Section 7, can raise the realized rate, which group-conditional calibration controls and which we recommend for deployment. We probe both edges in Section 7 (Table 12, Appendix N); the two mechanisms answer different questions (Figure 1).

Who runs the test. When the regulated firm develops or deploys its own model (a provider under Art. 3(3) of the AI Act), provider-side provenance discharges Article 50’s marking obligation directly. For an auditor who will not trust the log, Proposition 2 (Appendix B) shows a binding commitment plus an independently held calibration set restores the $\leq \alpha$ guarantee, though coverage carries none.

Feasibility. The infrastructure is modest: a desk’s annual model-drafted volume (10^8 to 10^9 tokens) fits compressed exact-match indices already serving trillion-token corpora at sub-second latency (Liu et al., 2024; Xu et al., 2025), over records SEC 17a-4 already requires the regulated firm to maintain (U.S. Securities and Exchange Commission, 2023).

Watermark	None	Light	DIPPER	GPT-4o	Rec. ²
Green-list	0.99	0.86	0.27	0.14	0.05
SynthID-Text	0.98	0.83	0.24	0.12	0.04
Aaronson–Gumbel	0.99	0.88	0.29	0.16	0.06
Unigram	0.98	0.92	0.39	0.24	0.11

Table 3: T1 (stateless, third-party-verifiable): coverage (TPR at each method’s nominal 1% threshold) under permitted edits at $L=100$ (500 positive and 500 negative examples per condition per seed, five seeds). All entries are means over five seeds; the largest standard deviation is 0.05. Realized FPR on the full third-party EDGAR null set is 1.1 to 1.6%. Every watermark collapses under paraphrase (Table 8, Appendix I); DetectGPT and GLTR cannot hold 1% (Appendix J).

Method	None	Light	DIPPER	GPT-4o	Rec. ²
Conformal retrieval	1.00	0.99	0.94	0.88	0.79

Table 4: T2 (provider-side, log-holding): coverage under the same conditions and sample sizes as Table 3. All entries are means over five seeds; the largest standard deviation is 0.03. Realized FPR is 1.0%. Under exchangeable calibration, the conformal retrieval test holds coverage at $\leq 1\%$ FPR (Proposition 1) at every length, sidestepping Theorem 1 because its detector reads the stored log rather than the text alone. T2 answers a different question from T1 (Figure 1).

7 Experiments

Two regimes, one scaling. The theorem and the experiments sit at different operating points, and conflating them would be a mistake. Theorem 1 bites in the high-quality regime $\gamma \approx 0.02$ that quality-sensitive deployments target but none of the four benchmarked schemes uses; there the horizon sits above sentence length and the triad is infeasible. The experiments run at deployed $\gamma \approx 0.19$ to 0.24 for the logit-shift schemes (Table 6), where the same formula puts the horizon below $L=100$, so the theorem does *not* forbid the triad and the empirical collapse is not its consequence; the only link is the shared $1/(\eta\gamma)$ scaling and Corollary 2, with edit-channel contraction, not the length floor, failing the schemes (Table 8, Figure 5). Because the feasible length exceeds L^* (Theorem 3), the fractions computed at L^* are a *lower* bound on the true uncoverable fraction.

Two threat models. We evaluate within the two models of Figure 1, not as one leaderboard: T1 (the four watermarks plus zero-shot DetectGPT and GLTR) is stateless and third-party-verifiable; T2 (our conformal test) is provider-side and log-

holding. T2 does not “beat” T1; under the triad on short text T1 is infeasible while T2 is feasible.

Setup. We evaluate on SEC EDGAR 10-K/10-Q filings and earnings-call transcripts. The null is text not produced by the host model; the host set is model-generated summaries of filing excerpts from OPT-6.7B (Zhang et al., 2022) (the primary host generator) and Llama-3 (Grattafiori et al., 2024) in a financial setting (Wu et al., 2023; Chen et al., 2021). All T1 watermark results use OPT-6.7B as the host generator. We watermark with green-list (Kirchenbauer et al., 2023), SynthID-Text (Dathathri et al., 2024), Aaronson–Gumbel (Aaronson, 2023), and Unigram (Zhao et al., 2024), plus zero-shot DetectGPT (Mitchell et al., 2023) and GLTR (Gehrmann et al., 2019) as T1 base-lines (Wu et al., 2025). Edits, in increasing aggressiveness, are disclaimer/redaction (*light*), DIPPER (Krishna et al., 2023), GPT-4o paraphrase (gpt-4o-2024-08-06) (OpenAI, 2024), and DIPPER twice (*recursive*). The construction uses a hybrid embedding-plus-n-gram index with split-conformal calibration ($n=2000$); we report TPR at each method’s nominal 1% threshold over five seeds, with realized false-positive rates, measured on unedited third-party text, in the table captions (watermarks realize 1.1 to 1.6%, slightly above T2’s 1.0%, so the comparison is conservative in the watermarks’ favor); each seed varies the random generation of host-model outputs and the test sample, with the conformal calibration set fixed across seeds; prompt variation is not sampled (a single host prompt and a single paraphrase prompt are used throughout, listed in Table 14).

Detection under permitted edits. Tables 3 and 4 report coverage at $L=100$ (above the median regulated unit of 26 tokens; Figure 4 extends to $L=25$). Every watermark collapses under paraphrase: Unigram reaches 0.39 under DIPPER, while the construction holds 0.94 at $\leq 1\%$ FPR. We do not credit this to the theorem: deployed $\gamma \approx 0.19$ to 0.24 puts the horizon below $L=100$, so Theorem 1 does *not* forbid the triad here. The optimal green-count test (Table 8, Appendix I) more than doubles DIPPER coverage ($0.27 \rightarrow 0.71$) yet still misses 0.95, so contraction binds even for the best detector; realized coverage below the resampling-model prediction (0.71 versus 0.97) indicates that $\hat{\eta}$ systematically overstates the operative contraction across all channels, including light edits (Table 8), suggesting a bias in the detection-statistic ratio rather than

a paraphrase-specific effect; the direction is conservative, implying higher true L^* and a larger uncovered fraction than reported. Coverage crosses 0.95 near $L \approx 500$ to 700 (Figure 4, Appendix H) while the construction is length-flat to $L=25$.

Calibration and ablation. Table 10 (Appendix L) reports false-positive calibration: watermark z -tests drift above the 1% level on repetitive boilerplate, whereas split-conformal meets the marginal target; on a matched hard-negative stratum, group-conditional calibration (Romano et al., 2020; Gibbs et al., 2025) restores per-stratum control (Table 12). The ablation (Appendix M) shows FPR control is invariant to the score, edit-robust coverage comes from semantic similarity, and a larger calibration set concentrates the realized rate around the target, with spread following $\text{Beta}(n+1-k, k)$. Varying the host generator from OPT-6.7B to GPT-4o lowers conformal coverage only from 0.94 to 0.88 at a held 1% FPR (Appendix O), so the advantage is not an artifact of an easily separable generator.

Takeaways. Watermark coverage becomes usable only past $L \approx 500$ to 700; at $L=100$ under paraphrase it stays below 0.4. Within the provider-side model, the construction is the only method whose false-positive rate is a guarantee rather than a measurement.

8 Discussion

When watermarking still suffices. The horizon constrains *short* outputs, and the operative reading is its contrapositive: once an output is long enough, a bounded-distortion watermark meets the triad on its own. The governing length is the operative L_{op} , not the lower bound L^* (Theorem 3), and under DIPPER paraphrase this crossover sits near 500 to 700 tokens (Figure 4); under recursive paraphrase the $1/\eta$ scaling of Table 6 places it near 1 100 to 1 600 tokens, comparable to filing length, so the division of labor is itself edit-strength dependent. There the watermark is the instrument of choice rather than a fallback, since it is stateless and third-party-verifiable where the provider log is not. The resulting policy is length- and unit-aware: watermark the long outputs, and route the short ones, where supervisory weight concentrates (Table 2), through the mandated log. This is the division of labor the European Code of Practice’s “no single technique” position anticipates (European

Commission, AI Office, 2026); our results give it a quantitative boundary, since the regulated unit rather than a vague notion of “enough text” is what decides which mechanism applies.

No parametric escape. Since $L^* \propto 1/(\eta\gamma)$, the only parametric way to shrink it is to raise γ or lower η , neither of which a quality-seeking scheme controls: γ must be small for deployability, and η is fixed by the permitted edits. The green-list interaction between distortion and contraction (Appendix E) shifts the per-pair ratio by only $O(\gamma)$. With the matching bound confirming $L_{\text{feasible}} = \Theta(L^*)$, the only exit leaves the token-perturbing class entirely, by changing the detection mechanism (Section 6).

9 Conclusion

We formalized regulated provenance as a triad and proved that no bounded-distortion watermark satisfies it below a closed-form horizon L^* , with a matching bound that makes the horizon order-tight in the distortion and contraction parameters rather than a loose necessary condition. The contraction factor entering the bound is a proven upper bound for the memoryless resampling channel and a calibrated estimate under realistic paraphrase, a distinction we keep explicit throughout: the barrier covers most sentence-level SEC EDGAR outputs by our estimates, while the empirical collapse of deployed watermarks reflects edit-channel contraction at their operating budgets rather than the length floor itself, the two regimes linked only by the shared $1/(\eta\gamma)$ scaling. A conformally calibrated retrieval log escapes the impossibility under a provider-side threat model, with a false-positive rate at or below the target at every length under exchangeable calibration. As marking obligations take effect, the triad is a design lens for provenance mechanisms, and the regulated unit, not raw output length, is what decides which mechanism can carry the mandate.

Limitations

Theorem 1 applies only to γ -bounded token-distribution-perturbing schemes under an edit channel with strictly positive contraction coefficient. Theorem 2 is proved for the token-resampling channel and the green-list family; we do not claim a matching upper bound for arbitrary channels. The bound L^* is a distribution-free necessary condition,

not an exact achievable length, so it is conservative. The proxy estimates $\hat{\eta}$ and $\hat{\gamma}$ are measured on our data: $\hat{\gamma}$ is a token-level mean matching Definition 3’s average, while $\hat{\eta}$ for the paraphrase channel is a ratio of detection-statistic divergences that need not upper-bound the per-pair contraction the theorem uses; hence the paraphrase horizon and EDGAR fraction are calibrated estimates rather than strict consequences of Theorem 1. The construction’s coverage depends on index quality, and its false-positive guarantee assumes exchangeability between calibration and deployment text; coverage further degrades against a fully adaptive editor. Empirically, we evaluate solely on financial text.

Ethical considerations

Our construction logs model outputs, raising data-retention and privacy questions. In regulated settings retention is already mandated, but other deployments should minimize logs, apply access controls, and enforce retention limits, for instance by storing only hashed or embedding representations. Provenance mechanisms can be misused to surveil or falsely attribute authorship; the false-positive guarantee we emphasize limits wrongful attribution, and we encourage any deployment to publish its calibration target. When coverage fails, genuine model outputs go unflagged, so deployments relying on provenance for oversight should weigh missed-detection risk alongside misattribution risk. False positives are not evenly distributed: the repetitive-boilerplate stratum draws higher rates, so the risk of misattribution falls disproportionately on writers whose style resembles model boilerplate; group-conditional calibration mitigates but does not eliminate this. All SEC EDGAR filings used are U.S. government public-domain records. Earnings-call transcripts were obtained from public sources and are excluded from the code release because redistribution rights are unconfirmed. OPT is released under a non-commercial license; Llama-3 under the Llama 3 Community License; GPT-4o outputs are governed by OpenAI’s Terms of Service. Table 12 reports an embedding-targeted adaptive editor that lowers coverage. EDGAR filings from 2020–2025 overlap with standard pretraining corpora; the host-generator sensitivity sweep (Appendix O) controls for this. Finally, a length-aware routing policy mandates retaining more short communications, carrying a privacy cost even where record-handling rules apply; deployments should

minimize retention where law allows.

Reproducibility

The released code is available at <https://github.com/phx0i/provenance-horizon>.

Acknowledgments

We thank the anonymous reviewers and area chair for their detailed feedback, which substantially improved the paper. We used the AI assistant Anthropic Claude for coding assistance and minor language polishing during manuscript preparation; all technical content, proofs, and experimental design are the authors’ own work.

References

- Scott Aaronson. 2023. Watermarking of large language models. Talk, “Large Language Models and Transformers” workshop, Simons Institute for the Theory of Computing, Berkeley, 17 August 2023.
- Anastasios N. Angelopoulos, Rina Foygel Barber, and Stephen Bates. 2024. Theoretical foundations of conformal prediction. *ArXiv:2411.11824*.
- Anastasios N. Angelopoulos and Stephen Bates. 2023. Conformal prediction: A gentle introduction. *Found. Trends Mach. Learn.*, 16(4):494–591.
- J. Bretagnolle and C. Huber. 1978. *Estimation des densités : Risque minimax*, pages 342–363. Springer Berlin Heidelberg.
- Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, et al. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.
- Zhongze Cai, Shang Liu, Hanzhao Wang, Huaiyang Zhong, and Xiaocheng Li. 2026. Towards better statistical understanding of watermarking LLMs. *Journal of the American Statistical Association*, 121(553):247–258.
- Souradip Chakraborty, Amrit S. Bedi, Sicheng Zhu, Bang An, Dinesh Manocha, and Furong Huang. 2024. Position: On the possibilities of AI-generated text detection. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*, Proceedings of Machine Learning Research, pages 6093–6115. PMLR.

- Zhiyu Chen, Wenhui Chen, Charese Smiley, Sameena Shah, Iana Borova, Dylan Langdon, Reema Moussa, Matt Beane, Ting-Hao Huang, Bryan Routledge, and William Yang Wang. 2021. FinQA: A dataset of numerical reasoning over financial data. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 3697–3711. Association for Computational Linguistics.
- Miranda Christ, Sam Gunn, and Or Zamir. 2024. Undetectable watermarks for language models. In *The Thirty Seventh Annual Conference on Learning Theory, June 30 - July 3, 2024, Edmonton, Canada*, Proceedings of Machine Learning Research, pages 1125–1139. PMLR.
- Coalition for Content Provenance and Authenticity. 2024. C2PA technical specification, version 2.1.
- Sumanth Dathathri, Abigail See, Sumedh Ghaisas, Posen Huang, Rob McAdam, Johannes Welbl, Vandana Bachani, Alex Kaskasoli, Robert Stanforth, Tatiana Matejovicova, Jamie Hayes, Nidhi Vyas, Majd Al Meray, Jonah Brown-Cohen, Rudy Bunel, Borja Balle, A. Taylan Cemgil, Zahra Ahmed, Kitty Stacpoole, et al. 2024. Scalable watermarking for identifying large language model outputs. *Nature*, 634(8035):818–823.
- Yanai Elazar, Akshita Bhagia, Ian Magnusson, Abhishek Ravichander, Dustin Schwenk, Alane Suhr, Evan Pete Walsh, Dirk Groeneveld, Luca Soldaini, Sameer Singh, Hannaneh Hajishirzi, Noah A. Smith, and Jesse Dodge. 2024. What’s in my big data? In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.
- European Commission. 2026. Guidelines on Transparency Obligations for Providers and Deployers of Certain AI Systems (Article 50 of regulation (EU) 2024/1689). Adopted 20 July 2026.
- European Commission, AI Office. 2026. Code of practice on transparency of AI-generated content, section 1: Providers. Published 10 June 2026. Found adequate by the Commission and AI Board, 8–9 July 2026.
- European Parliament and Council of the European Union. 2024. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on Artificial Intelligence (Artificial Intelligence Act), Article 50. Official Journal of the European Union, OJ L 2024/1689, 12.7.2024.
- European Parliament and Council of the European Union. 2026. Regulation (EU) 2026/1744 of the European Parliament and of the Council of 8 July 2026 amending regulations (EU) 2024/1689, (EU) 2018/1139 and (EU) 2023/1230 (Digital Omnibus on AI). Official Journal of the European Union, OJ L 2026/1744, 24.7.2026; in force 27.7.2026. Grants a machine-readable marking transition to 2 December 2026 for systems placed on the market before 2 August 2026.
- Jaiden Fairoze, Sanjam Garg, Somesh Jha, Saeed Mahlouljifar, Mohammad Mahmood, and Mingyuan Wang. 2025. Publicly-detectable watermarking for language models. *IACR Commun. Cryptol.*, 1(4).
- Pierre Fernandez, Antoine Chaffin, Karim Tit, Vivien Chappelier, and Teddy Furon. 2023. Three bricks to consolidate watermarks for large language models. In *IEEE International Workshop on Information Forensics and Security, WIFS 2023, Nürnberg, Germany, December 4-7, 2023*, pages 1–6. IEEE.
- Paolo Ferragina and Giovanni Manzini. 2000. Opportunistic data structures with applications. In *41st Annual Symposium on Foundations of Computer Science, FOCS 2000, Redondo Beach, California, USA, November 12-14, 2000*, pages 390–398. IEEE Computer Society.
- Financial Industry Regulatory Authority. 2014. FINRA Rule 3110: Supervision.
- Sebastian Gehrmann, Hendrik Strobelt, and Alexander M. Rush. 2019. GLTR: statistical detection and visualization of generated text. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28 - August 2, 2019, Volume 3: System Demonstrations*, pages 111–116. Association for Computational Linguistics.
- Isaac Gibbs, John J. Cherian, and Emmanuel J. Candès. 2025. Conformal prediction with conditional guarantees. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 87(4):1100–1126.
- Noah Golowich and Ankur Moitra. 2024. Edit distance robust watermarks via indexing pseudorandom codes. In *Advances in Neural Information Processing Systems 37: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10-15, 2024*.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, et al. 2024. The Llama 3 herd of models. ArXiv:2407.21783.
- Hengzhi He, Shirong Xu, Alexander Nemecek, Jiping Li, Erman Ayday, and Guang Cheng. 2026. Optimal watermark generation under type I and type II errors. *Biometrika*. Advance article.
- Zhengmian Hu, Lichang Chen, Xidong Wu, Yihan Wu, Hongyang Zhang, and Heng Huang. 2024. Unbiased watermark for large language models. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

- Baihe Huang, Hanlin Zhu, Banghua Zhu, Kannan Ramchandran, Michael I. Jordan, Jason D. Lee, and Jiantao Jiao. 2023. Towards optimal statistical watermarking. In *Socially Responsible Language Modelling Research (SoLaR), NeurIPS 2023 Workshop*. Non-archival workshop paper.
- Nikola Jovanović, Robin Staab, and Martin T. Vechev. 2024. Watermark stealing in large language models. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*, Proceedings of Machine Learning Research, pages 22570–22593. PMLR.
- John Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, and Tom Goldstein. 2023. A watermark for large language models. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, Proceedings of Machine Learning Research, pages 17061–17084. PMLR.
- John Kirchenbauer, Jonas Geiping, Yuxin Wen, Manli Shu, Khalid Saifullah, Kezhi Kong, Kasun Fernando, Aniruddha Saha, Micah Goldblum, and Tom Goldstein. 2024. On the reliability of watermarks for large language models. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.
- Kalpesh Krishna, Yixiao Song, Marzena Karpinska, John Wieting, and Mohit Iyyer. 2023. Paraphrasing evades detectors of AI-generated text, but retrieval is an effective defense. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.
- Rohith Kuditipudi, John Thickstun, Tatsunori Hashimoto, and Percy Liang. 2024. Robust distortion-free watermarks for language models. *Trans. Mach. Learn. Res.*, 2024.
- Lucien Le Cam. 1986. *Asymptotic Methods in Statistical Decision Theory*. Springer Series in Statistics. Springer-Verlag, New York.
- Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J. Tibshirani, and Larry Wasserman. 2018. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523):1094–1111.
- Xiang Li, Feng Ruan, Huiyuan Wang, Qi Long, and Weijie J. Su. 2025. A statistical framework of watermarks for large language models: Pivot, detection efficiency and optimal rules. *The Annals of Statistics*, 53(1):322–351.
- Xiang Li, Feng Ruan, Huiyuan Wang, Qi Long, and Weijie J. Su. 2026. Robust detection of watermarks for large language models under human edits. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 88(2):491–515.
- Jiacheng Liu, Taylor Blanton, Yanai Elazar, Sewon Min, Yen-Sung Chen, Arnavi Chheda-Kothary, Huy Tran, Byron Bischoff, Eric Marsh, Michael Schmitz, Cassidy Trier, Aaron Sarnat, Jenna James, Jon Borchardt, Bailey Kuehl, Evie Yu-Yen Cheng, Karen Farley, Taira Anderson, David Albright, et al. 2025. OL-MoTrace: Tracing language model outputs back to trillions of training tokens. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 178–188. Association for Computational Linguistics.
- Jiacheng Liu, Sewon Min, Luke Zettlemoyer, Yejin Choi, and Hannaneh Hajishirzi. 2024. Infini-gram: Scaling unbounded n-gram language models to a trillion tokens. In *First Conference on Language Modeling*.
- Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D. Manning, and Chelsea Finn. 2023. DetectGPT: Zero-shot machine-generated text detection using probability curvature. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, Proceedings of Machine Learning Research, pages 24950–24962. PMLR.
- Jerzy Neyman and Egon Sharpe Pearson. 1933. On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London, Series A: Containing Papers of a Mathematical or Physical Character*, 231(694-706):289–337.
- OpenAI. 2024. GPT-4o system card.
- Julien Piet, Chawin Sitawarin, Vivian Fang, Norman Mu, and David Wagner. 2025. Mark my words: Analyzing and evaluating language model watermarks. In *IEEE Conference on Secure and Trustworthy Machine Learning, SaTML 2025*, pages 68–91. IEEE.
- Yury Polyanskiy and Yihong Wu. 2025. *Information Theory: From Coding to Learning*. Cambridge University Press.
- Yaniv Romano, Rina Foygel Barber, Chiara Sabatti, and Emmanuel J. Candès. 2020. With Malice Toward None: Assessing Uncertainty via Equalized Coverage. *Harvard Data Science Review*, 2(2).
- Vinu Sankar Sadasivan, Aounon Kumar, Sriram Balasubramanian, Wenxiao Wang, and Soheil Feizi. 2025. Can AI-generated text be reliably detected? stress testing AI text detectors under various attacks. *Trans. Mach. Learn. Res.*, 2025.
- Glenn Shafer and Vladimir Vovk. 2008. A tutorial on conformal prediction. *J. Mach. Learn. Res.*, 9:371–421.
- Alexandre B. Tsybakov. 2009. *Introduction to Non-parametric Estimation*. Springer series in statistics. Springer.

U.S. Securities and Exchange Commission. 2023. Records to Be Preserved by Certain Exchange Members, Brokers and Dealers, 17 CFR § 240.17a-4. Code of Federal Regulations.

Vladimir Vovk, Alexander Gammernan, and Glenn Shafer. 2005. *Algorithmic Learning in a Random World*. Springer-Verlag, New York.

Junchao Wu, Shu Yang, Runzhe Zhan, Yulin Yuan, Lidia S. Chao, and Derek Fai Wong. 2025. A survey on LLM-generated text detection: Necessity, methods, and future directions. *Comput. Linguistics*, 51(1):275–338.

Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kam-badur, David Rosenberg, and Gideon Mann. 2023. BloombergGPT: A large language model for finance. ArXiv:2303.17564.

Yangxinyu Xie, Xiang Li, Tanwi Mallick, Weijie J. Su, and Ruixun Zhang. 2025. Debiasing watermarks for large language models via maximal coupling. *Journal of the American Statistical Association*, 120(551):1424–1436.

Hao Xu, Jiacheng Liu, Yejin Choi, Noah A. Smith, and Hannaneh Hajishirzi. 2025. Infini-gram mini: Exact n-gram search at the internet scale with FM-index. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, EMNLP 2025, Suzhou, China, November 4-9, 2025*, pages 24944–24969. Association for Computational Linguistics.

Hanlin Zhang, Benjamin L. Edelman, Danilo Francati, Daniele Venturi, Giuseppe Ateniese, and Boaz Barak. 2024. Watermarks in the sand: Impossibility of strong watermarking for language models. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*, Proceedings of Machine Learning Research, pages 58851–58880. PMLR.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. 2022. OPT: Open pre-trained transformer language models. ArXiv:2205.01068.

Xuandong Zhao, Prabhanjan Ananth, Lei Li, and Yu-Xiang Wang. 2024. Provable robust watermarking for AI-generated text. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

A Construction Algorithm

Algorithm 1 specifies the conformal retrieval-log provenance test of Section 6.

Algorithm 1: Conformal retrieval-log provenance

Input: output log $\mathcal{L}$; calibration set

$\{y_i\}_{i=1}^n$; target FPR α ; query z

Output: provenance decision in $\{0, 1\}$ with marginal FPR $\leq \alpha$

```

1 for  $i \leftarrow 1$  to  $n$  do
2    $s_i \leftarrow \max_{r \in \mathcal{L}} \text{sim}(y_i, r)$ ;
3 end
4  $\hat{\tau} \leftarrow$  the  $\lceil (1 - \alpha)(n + 1) \rceil$ -th smallest of
    $\{s_i\}$ ;
5  $s(z) \leftarrow \max_{r \in \mathcal{L}} \text{sim}(z, r)$ ;
6 return  $1[s(z) > \hat{\tau}]$ ;
```

B Auditor Guarantee under Commitment

Proposition 1 assumes the verifier trusts the provider’s log. The contested case is an auditor who does not; two ingredients restore the guarantee.

Proposition 2 (Auditor-verifiable calibration under commitment). *Suppose the provider publishes a binding commitment to the log $\mathcal{L}$ and to the similarity function $\text{sim}(\cdot, \cdot)$ (including model identity and index configuration) before any query is adjudicated, and the calibration set $\{y_i\}$ is held independently of the provider and is exchangeable with deployment third-party text. Then for any fresh third-party Y exchangeable with the calibration set, $\Pr[s(Y) > \hat{\tau}] \leq \alpha$ holds even for a verifier who does not trust the provider, at every length L and under every edit applied to the provider’s own outputs.*

Proof. The commitment fixes both $\mathcal{L}$ and $\text{sim}(\cdot, \cdot)$, hence $s(\cdot)$ is determined before Y is seen; $s(Y)$ and $\{s(y_i)\}$ are exchangeable and the rank of $s(Y)$ is uniform on $\{1, \dots, n + 1\}$, giving $\Pr[s(Y) > \hat{\tau}] \leq \alpha$ as in Proposition 1. Holding $\{y_i\}$ independently of the provider removes any influence on $\hat{\tau}$. $\square$

C Positioning Detail

Work	FPR bound	Edit rob.	Cov. floor	Reg. grd.	Result type
Kirchenbauer et al. (2023)	emp.	part.	no	no	scheme
Kuditipudi et al. (2024)	exact p	part.	no	no	scheme
Dathathri et al. (2024)	emp.	part.	no	no	scheme
Zhang et al. (2024)	no	unbd.	no	no	impos.
Sadasivan et al. (2025)	TV	recur.	no	no	impos.
Huang et al. (2023)	Type I	part.	tok.	no	opt. r.
This work	$\leq \alpha$	ch. η	$\geq 1-\beta$	yes	all 3

Table 5: Positioning against nearest prior work.

D Full Proof of Theorem 1

We restate and complete the argument. Fix a γ -bounded scheme and an edit channel $\mathcal{E}$ with KL contraction coefficient $\eta \in (0, 1)$ (Definition 4). No invariance of Q under $\mathcal{E}$ is assumed.

Step 1: testing reduction. By Definitions 1 and 2 the detector, acting on workflow-processed text, is a (possibly randomized) test for $H_0 : Z \sim \mathcal{E} \circ Q$ versus $H_1 : Z \sim \mathcal{E} \circ P_k$ with Type-I error at most α and Type-II error at most β . The Le Cam two-point inequality (Le Cam, 1986; Tsybakov, 2009) states that for any test the sum of the two errors is at least $1 - \text{TV}(P_0, P_1)$; hence (2).

Step 2: contraction. The Bretagnolle–Huber inequality (Bretagnolle and Huber, 1978) bounds $\text{TV}(P, R) \leq \sqrt{1 - e^{-\text{KL}(R\|P)}}$ (we use this $\sqrt{1 - e^{-\text{KL}}}$ form, not the $\frac{1}{2}e^{-\text{KL}}$ variant, since the constant propagates into L^*). The strong data-processing inequality with coefficient η (Polyanskiy and Wu, 2025) gives $\text{KL}(\mathcal{E} \circ P_k \| \mathcal{E} \circ Q) \leq \eta \text{KL}(P_k \| Q)$ for the two laws P_k and Q directly. Composing yields (3). (TV is symmetric; we use $\text{KL}(P_k \| Q)$ to match Definition 3.) This is where the prior reliance on $\mathcal{E} \circ Q = Q$ is removed: applying the contraction to the pair (P_k, Q) needs only that η is the channel’s contraction coefficient.

Step 3: distortion budget. For autoregressive laws the KL chain rule gives $\text{KL}(P_k \| Q) = \sum_{t=1}^L \mathbb{E}_{x_{<t} \sim P_k} \text{KL}(P_k(\cdot | x_{<t}) \| Q(\cdot | x_{<t})) \leq L\gamma$ by Definition 3, with the expectation taken under the watermarked path measure P_k as the chain rule requires.

Step 4: combine. Substituting into (3) and chaining with (2): $1 - \alpha - \beta \leq \sqrt{1 - e^{-\eta L \gamma}}$. Squaring and rearranging, $e^{-\eta L \gamma} \leq 1 - (1 - \alpha - \beta)^2$, hence $\eta L \gamma \geq \ln \frac{1}{1 - (1 - \alpha - \beta)^2}$ and $L \geq L^*$. The right side is finite and positive whenever $0 < \alpha + \beta < 1$.

E Proof of Theorem 2

Let the green-list scheme have green fraction $g = \frac{1}{2}$ and per-token green probability $p = g + \delta$ under P_k , with $\delta > 0$ the bias. The per-token KL between the green indicators is $f(\delta) := \text{KL}(\text{Bern}(g+\delta) \| \text{Bern}(g))$, the exact Bernoulli divergence. For any budget $\gamma > 0$, the budget-consistent bias $\delta^*(\gamma)$ is the unique positive solution of $f(\delta^*) = \gamma$, obtained by inverting the exact KL. Since $f(\delta) = \frac{\delta^2}{2g(1-g)} + O(\delta^4)$, the leading-order inverse is $\delta \approx \sqrt{2g(1-g)\gamma}$, at $g = \frac{1}{2}$, the exact budget-consistent bias satisfies $\delta^{*2} = (\gamma/2)(1 - \gamma/3 + O(\gamma^2))$, making the true sufficient length larger than the leading-order formula by a factor $1/(1 - \gamma/3)$: a 0.67% correction at the aspirational $\gamma=0.02$ and under 10% at the largest deployed budget $\gamma=0.24$. The displayed formulas use the leading-order equivalence $\delta^{*2} \approx \gamma/2$ for readability; the $\Theta(1/(\eta\gamma))$ scaling is exact regardless.

Under the token-resampling channel R_ρ , each token is redrawn from Q with probability ρ , so the post-edit green rate is $p_e = g + (1 - \rho)\delta$ under H_1 and g under H_0 . The detection statistic is the green count $S = \sum_{t=1}^L \mathbb{1}[\text{token } t \text{ green}]$. Set the threshold at the midpoint $\tau = L(g + \frac{(1-\rho)\delta}{2})$.

By Azuma–Hoeffding for bounded martingale differences, with the gap $(1 - \rho)\delta/2$ between τ/L and each mean, the Type-I error is

$$\Pr_{H_0}[S \geq \tau] \leq \exp\left(-\frac{1}{2}L(1 - \rho)^2\delta^2\right),$$

and symmetrically the Type-II error $\Pr_{H_1}[S < \tau]$ obeys the same bound. Writing $\eta = (1 - \rho)^2$ and using the exact budget-consistent bias δ^* satisfying $f(\delta^*) = \gamma$, both error exponents equal $L\eta\delta^{*2}/2$. The sufficient condition $L\eta\delta^{*2}/2 \geq \ln(1/\min(\alpha, \beta))$ yields

$$L \geq \frac{2}{\eta\delta^{*2}(\gamma)} \ln \frac{1}{\min(\alpha, \beta)},$$

which is (4).

To combine with Theorem 1, note that under the green-list idealization both P_k and Q factorize over green indicators, so $\text{KL}(P_k \| Q) = Lf(\delta)$ and $\text{KL}(\mathcal{E} \circ P_k \| \mathcal{E} \circ Q) = Lf(c\delta)$ with $c = 1 - \rho$ and the per-pair contraction ratio $f(c\delta)/f(\delta)$. At $g = \frac{1}{2}$,

$$f(\epsilon) = \sum_{j \geq 1} a_j \epsilon^{2j}, \quad a_j = \frac{2^{2j-1}}{j(2j-1)} > 0,$$

the even series of $\frac{1}{2}[(1+2\epsilon)\ln(1+2\epsilon) + (1-2\epsilon)\ln(1-2\epsilon)]$, with $a_1=2$, $a_2=\frac{4}{3}$, and every coefficient positive. Term by term,

$$f(c\delta) - c^2 f(\delta) = \sum_{j \geq 1} a_j \delta^{2j} (c^{2j} - c^2),$$

where the $j=1$ term vanishes and each $j \geq 2$ term is strictly negative for $c \in (0, 1)$. Hence $f(c\delta) \leq c^2 f(\delta)$ for every δ , so $\eta = (1 - \rho)^2$ is a valid upper bound on the per-pair contraction. Both bounds are $\Theta(1/((1 - \rho)^2 \gamma))$ and

$$\frac{L_{\text{ach}}}{L^*} = \frac{2\gamma \ln(1/\min(\alpha, \beta))}{\delta^{*2}(\gamma) \ln \frac{1}{1-(1-\alpha-\beta)^2}},$$

which equals $4 \ln(1/\min(\alpha, \beta)) / \ln \frac{1}{1-(1-\alpha-\beta)^2}$ at leading order in γ (about 8.6 at $(\alpha, \beta) = (10^{-2}, 0.05)$). The exact binomial minimum L_{exact} runs 3.8–4.1 times L^* (Table 7), so $L_{\text{feasible}} = \Theta(L^*)$.

Sharp constant via a balanced threshold.

Theorem 3 (Operative feasible length). *For the channel R_ρ and the green-list family of Theorem 2, the green-count test whose threshold (set using ρ) divides the post-edit gap in proportion $\sqrt{\ln(1/\alpha)} : \sqrt{\ln(1/\beta)}$ achieves α -calibration and $(1 - \beta)$ -coverage for every $L \geq L_{\text{op}}$, with $\eta = (1 - \rho)^2$, where*

$$L_{\text{op}} = \frac{(\sqrt{\ln(1/\alpha)} + \sqrt{\ln(1/\beta)})^2}{2\eta \delta^{*2}(\gamma)}. \quad (5)$$

Because $(\sqrt{a} + \sqrt{b})^2 \leq 4 \max(a, b)$ (with $a = \ln \frac{1}{\alpha}$, $b = \ln \frac{1}{\beta}$), this strictly improves the midpoint bound (4) whenever $\alpha \neq \beta$, and with Theorem 1 gives $L^* \leq L_{\text{feasible}} \leq L_{\text{op}}$.

Proof. The midpoint threshold of (4) charges both error types the worse budget $\min(\alpha, \beta)$, which is the only reason its constant is loose. Place the threshold instead at rate $g + x\Delta$ with $\Delta = (1 - \rho)\delta$ the post-edit gap and $x \in (0, 1)$. Hoeffding on the green count gives Type-I error $\Pr_{H_0}[S \geq L(g + x\Delta)] \leq \exp(-2Lx^2\Delta^2)$ and, symmetrically, Type-II error $\leq \exp(-2L(1 - x)^2\Delta^2)$. Requiring the first $\leq \alpha$ and the second $\leq \beta$ gives $x\Delta \geq \sqrt{\ln(1/\alpha)/(2L)}$ and $(1 - x)\Delta \geq \sqrt{\ln(1/\beta)/(2L)}$; adding the two,

$$\Delta \geq \frac{\sqrt{\ln(1/\alpha)} + \sqrt{\ln(1/\beta)}}{\sqrt{2L}},$$

that is, $L \geq (\sqrt{\ln(1/\alpha)} + \sqrt{\ln(1/\beta)})^2 / (2\Delta^2)$, with the minimizing split $x = \sqrt{\ln(1/\alpha)} / (\sqrt{\ln(1/\alpha)} + \sqrt{\ln(1/\beta)})$. Using $\Delta^2 = \eta \delta^{*2}$ at $g = \frac{1}{2}$ yields (5). Numerically $L_{\text{op}}/L^* \approx 7.0$ and $L_{\text{CLT}}/L^* \approx 3.7$ at $(\alpha, \beta, \gamma) = (10^{-2}, 0.05, 0.02)$. $\square$

F Estimating η and γ

We estimate γ as the mean per-token $\text{KL}(P_k(\cdot | x_{<t}) \| Q(\cdot | x_{<t}))$ at each scheme’s deployed strength, and η as the ratio $\widehat{\text{KL}}(\mathcal{E} \circ P_k \| Q) / \widehat{\text{KL}}(P_k \| Q)$ of detection-statistic KL after versus before editing, where $\widehat{\text{KL}}$ is the divergence between Gaussian fits to the detector’s z -statistic under the two hypotheses. Values are in Table 6; $\hat{\gamma}$ matches Definition 3’s average, while $\hat{\eta}$ is a detection-statistic ratio rather than the sequence-level per-pair contraction the theorem uses.

Watermark	γ (nats)	Edit channel	η
Green-list	0.19	light (redact+discl.)	0.83
SynthID-Text	0.15 [†]	DIPPER paraphrase	0.47
Aaronson–Gumbel	0.17 [†]	GPT-4o paraphrase	0.34
Unigram	0.24	recursive ($\times 2$)	0.21
quality budget (target) 0.02		no edit (identity) n/a	

Table 6: Estimated γ (per-watermark) and η (per-channel, on green-list). [†]Key-averaged; the per-step conditional γ of Definition 3 is larger for sampling-rule schemes.

G Scaling Agreement Detail

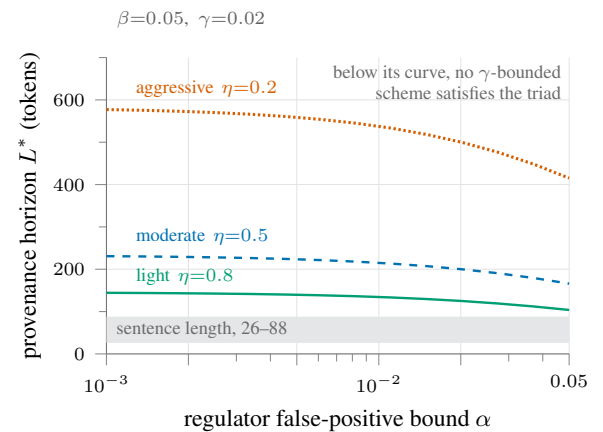


Figure 2: Provenance horizon L^* vs. α for three edit strengths ($\beta=0.05$, $\gamma=0.02$). Shaded band: sentence-length range (median 26 to P90 88 tokens).

edit setting	γ	η	L^*	L_{CLT}	L_{exact}
light	0.32	0.61	11	39	42
moderate	0.32	0.23	29	114	116
aggressive	0.32	0.04	168	687	697
high-quality WM	0.16	0.23	58	220	224

Table 7: Scaling agreement at $\alpha=0.01$, $\beta=0.05$ for the token-resampling probe. L^* : lower bound (1); L_{CLT} : Gaussian green-count threshold (two-variance); L_{exact} : exact binomial minimum.

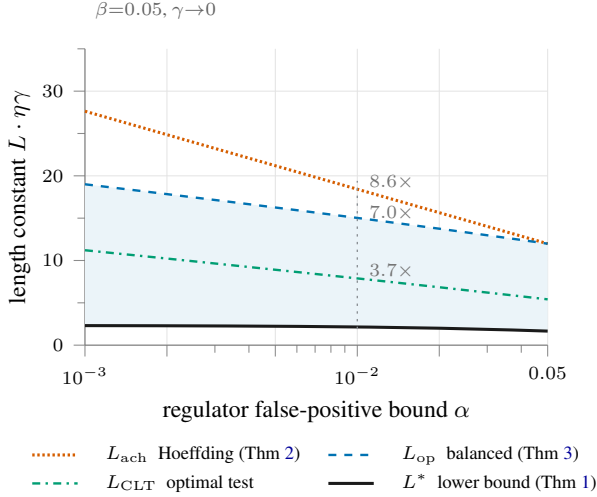


Figure 3: Feasible length in units of $1/(\eta\gamma)$ ($\gamma \rightarrow 0$ limit) vs. α at $\beta=0.05$. Shaded region: $L^* \leq L_{\text{feasible}} \leq L_{\text{op}}$.

H Coverage versus Output Length

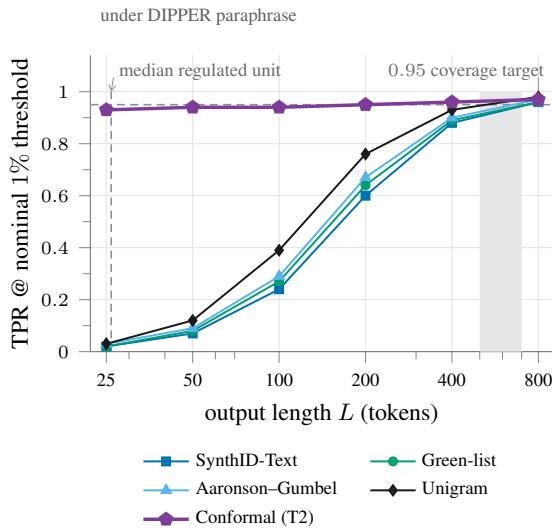


Figure 4: Coverage (TPR at nominal 1% threshold) vs. output length under DIPPER paraphrase. Dotted line: 0.95 coverage target.

I Optimal Detector at $L = 100$

Table 8 isolates detector suboptimality from edit-channel contraction at $L=100$, $\gamma=0.19$: the deployed z -test vs. the optimal Neyman–Pearson green-count test (exact level- α binomial), both at $\alpha=0.01$. The model column is the predicted coverage if $\hat{\eta}$ were the true per-pair contraction.

Edit channel	η	Deployed	Optimal	Model
light	0.83	0.86	0.97	1.00
DIPPER	0.47	0.27	0.71	0.97
GPT-4o	0.34	0.14	0.52	0.89
recursive	0.21	0.05	0.26	0.66

Table 8: Green-list coverage at $L=100$, $\gamma=0.19$: deployed z -test, optimal Neyman–Pearson green-count test, and resampling-model prediction.

J Zero-Shot Detectors

Edit	DetectGPT	GLTR
no edit	0.58 ± 0.05	0.49 ± 0.06
light	0.41 ± 0.05	0.34 ± 0.05
DIPPER	0.07 ± 0.03	0.06 ± 0.02
GPT-4o	0.04 ± 0.02	0.03 ± 0.02
recursive	0.02 ± 0.01	0.02 ± 0.01
realized FPR	4.8%	5.6%

Table 9: Zero-shot detector coverage at $L=100$ (mean $\pm$ std, five seeds) with realized FPR (Mitchell et al., 2023; Gehrmann et al., 2019).

K EDGAR Output Lengths

We extract risk-factor and MD&A sections of SEC EDGAR 10-K/10-Q filings and segment into sentences. The 1.2×10^6 sentences have median 26 tokens ($P_{90} = 88$, $P_{95} = 131$, $P_{99} = 298$). Filing-level excerpts have median ~ 1400 and mean ~ 5600 tokens.

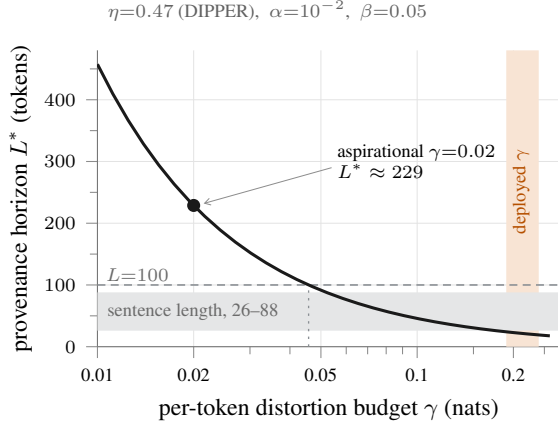


Figure 5: Provenance horizon vs. distortion budget γ at estimated DIPPER $\eta=0.47$ and $(\alpha, \beta)=(10^{-2}, 0.05)$. Shaded band: deployed γ range. Dashed line: $L=100$.

L False-Positive Calibration

Method	Marginal	Boilerplate
Green-list	$1.31 \pm 0.18\%$	2.9%
SynthID-Text	$1.12 \pm 0.15\%$	2.3%
Aaronson-Gumbel	$1.24 \pm 0.16\%$	2.6%
Unigram	$1.58 \pm 0.21\%$	3.4%
Ours (split-conformal)	$1.00 \pm 0.09\%$	1.4%

Table 10: False-positive rate on third-party text (target $\alpha=1\%$; mean $\pm$ std, five seeds), marginal and on repetitive boilerplate.

M Construction Ablation

Variant	TPR@1%FPR	Realiz. FPR
exact n-gram	0.31	1.0%
normalized LCS	0.55	1.0%
E5-large embedding	0.86	1.0%
hybrid (embed + n-gram)	0.88	1.0%
calibration $n=100$	0.87	$1.6 \pm 1.0\%$
calibration $n=500$	0.88	$1.2 \pm 0.4\%$
calibration $n=2000$	0.88	$1.0 \pm 0.2\%$

Table 11: Retrieval-score ablation at $L=100$ under GPT-4o paraphrase, with calibration-size sensitivity.

N Hard Negatives and Adaptive Editor

Table 12 reports FPR on hard negatives (topically matched third-party text from Llama-3-8B and earnings-call transcripts) under marginal and group-conditional calibration, and coverage under an embedding-targeted adaptive editor (best-of-eight DIPPER paraphrases by lowest E5 cosine).

False-positive rate (target 1%)	FPR
standard third-party null (marginal)	1.0%
hard-negative null (marginal)	2.3%
hard-negative null (group-conditional)	1.1%
Coverage under adaptive editor ($L=100$)	TPR
non-adaptive DIPPER (reference)	0.94
embedding-targeted, embedding index	0.61
embedding-targeted, hybrid index	0.68

Table 12: Hard-negative FPR (top) and adaptive-editor coverage (bottom) at $L=100$.

O Host-Set Sensitivity

Table 13 varies the host generator across four model scales, reporting conformal coverage at $L=100$ under DIPPER. Coverage declines gently from 0.94 to 0.88 while FPR holds at 1%.

Host generator	TPR@1%FPR	Realiz. FPR
OPT-6.7B (distinctive)	0.94	1.0%
Llama-3-8B	0.92	1.0%
Llama-3-70B	0.90	1.0%
GPT-4o (near null)	0.88	1.0%

Table 13: Conformal coverage at $L=100$ under DIPPER across host generators.

P Experimental Configuration

Component	Configuration
<i>Host generators</i>	
Primary	OPT-6.7B (facebook/opt-6.7b)
Scale sweep	Llama-3-8B, -70B ($T=1.0$, $\text{top-}p=0.95$); gpt-4o-2024-08-06,
Decoding	$T=1.0$, $\text{top-}p=0.95$ (watermarked); $T=0.7$, $\text{top-}p=0.9$ (GPT-4o host)
Max tokens	matched to target L per condition
<i>Watermark settings</i>	
Green-list	$g=0.5$, context width 1, logit bias 2.0
SynthID-Text	tournament depth 3
Gumbel	default key derivation
Unigram	$g=0.5$, context-free green partition
<i>Edit channels</i>	
Light	disclaimer prepend + PII redaction
DIPPER	kalpeshk2011/ dipper-paraphraser-xxl, default rev., lex 60, order 60
GPT-4o	gpt-4o-2024-08-06, $T=1.0$
Recursive	DIPPER applied twice
<i>Retrieval index</i>	
Embedding	E5-large-v2 (intfloat/e5-large-v2, default rev.)
N-gram	5-gram, BM25 weighting
Hybrid	$0.7 \times \text{embed} + 0.3 \times \text{n-gram}$
Search	exact (brute-force)

Table 14: Experimental configuration. GPT-4o paraphrase prompt: “Paraphrase the following text, preserving its meaning but changing the wording substantially.” Host-generation prompt: “Summarize the following SEC filing excerpt in 2-3 sentences, focusing on key risk factors and financial metrics.”